A Review on Use of Data Science for Visualization and Prediction of the COVID-19 Pandemic and Early Diagnosis of COVID-19 Using Machine Learning Models

Shiv Kumar Choubey and Harshit Naman

Abstract On December 30, 2019, the WHO China office was informed of a pneumonia-like disease with unknown etiology, from the Wuhan city of China. This disease was found to be caused by a new type of coronavirus. The virus was named severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2) and the disease caused by it was named as COVID-19. On March 11, the WHO declared COVID-19 a pandemic. The testing for COVID-19 disease can be broadly classified into two main techniques, firstly, by testing the patient's blood for immunoassays and second by PCR. The above two techniques are quite costly. Due to this, large-scale testing in developing countries like India is not practically possible. The novel coronavirus is highly infectious, and it spreads from one person to another even before the symptoms have appeared. So, the early detection of the virus will be a great way to stop this global pandemic from causing any more devastation and controlling its spread. In this paper, we review the role of technologies like artificial intelligence and deep learning in early detection, diagnosis, analysis (cure), and socio-economic impact of COVID-19. The purpose of this review paper is to provide a concise but judicious source of information to look over all the possible solutions. *Technologies used:* Artificial neural networks—Artificial neural networks are based on human brain and nervous system. An artificial neural network consists of several neurons and an activation function. ANNs have been used in diagnosis and early detection of several diseases like dengue and pneumonia. The same can be done in the case of COVID-19 by training the algorithm with suitable datasets. Deep Learning—Deep learning is a subclass of machine learning consisting of algorithms that are based on artificial neural networks. Deep learning is a very efficient way to handle large amount of data. Python libraries like Tensorflow are used in Linux-based systems for executing deep learning algorithms. *Visualization of the Pandemic* Several dashboards emerged gradually providing a global overview of the pandemic. Some of these are Upcode

S. K. Choubey (✉) · H. Naman
Department of Electronics and Communication Engineering, Birla Institute of Technology Mesra, Patna Campus 800014, India
e-mail: skchoubey@bitmesra.ac.in

H. Naman
e-mail: be15253.18@bitmesra.ac.in

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2020

C. Chakraborty et al. (eds.), *Internet of Medical Things for Smart Healthcare*, Studies in Big Data 80, https://doi.org/10.1007/978-981-15-8097-0_10

and NextStrain. Technologies like Python, Excel, R, and Tableau are used here for extracting data and visualizing them in the form of graphs and tables for the general public to understand. *Early Detection and Diagnosis* Artificial neural networks and deep learning can be used for early detection and diagnosis of the disease. The X-rays and CT images of the patients can be used as datasets. As of now the number of patients of COVID-19 in the world has increased to more than half a million, thus the dataset for training the neural networks and algorithms is quite large. This situation can be capitalized to make highly accurate neural networks using deep learning algorithms. The data can be extracted from press releases on the Internet and government databases by Web scraping. Libraries like Tensorflow can be used in training the models. *Tracking and Prediction* Artificial intelligence can be used to track and predict the spread of the coronavirus pandemic. We will try to throw some light on the past works in the area of epidemic prediction using AI. In 2015, neural networks were made for prediction of the Zika virus pandemic. These neural networks need to be trained again in accordance with the datasets of COVID-19. For example, Carnegie Mellon University algorithms used for predicting seasonal flu are being retrained with the datasets of COVID-19.

Keywords Coronavirus · COVID-19 · Data visualization · Artificial neural network · Deep learning · Pneumonia · Chest X-ray · CT images · AutoML · Artificial intelligence

1 Introduction

On December 30, 2019, the WHO China office was informed of pneumonia-like disease with unknown etiology, from the Wuhan city in Hubei province of China. On January 7, 2020, the Chinese authorities isolated the virus and notified the world about a new type of coronavirus. This virus has been named severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2), and its origins were identified to be a wet market in Wuhan city. The disease spread to the nearby countries of Japan and the Republic of Korea. On March 11, the WHO declared COVID-19 a pandemic as it spread across the world at an unimaginable rate, paralyzing the healthcare systems around the world. The epicenter of the pandemic shifted from Wuhan China to Italy and then to the USA [1].

We can classify testing for COVID-19 into two main techniques, firstly, by testing the patient's blood for immunoassays or coronavirus-related antibodies and proteins, and the second method is by identifying the genetic code of the virus. In the second process, the genetic material of the suspicious content is replicated by reverse transcription-polymerase chain reaction or RT-PCR. The above two techniques are quite cumbersome as well as costly [2]. Due to this, extensive scale testing in developing countries like India is not practically possible. As of now, ₹4500 is the estimated cost for COVID-19 tests in India. India has a population of more than 1.3 billion people, which makes testing even more costly [3].

So the question that arises is how helpful is extensive scale testing in eradicating this infection. The answer lies in the nature and behavior of this disease. The novel coronavirus is highly infectious; it spreads from one person to another even before the symptoms have appeared; this makes it more dangerous. So, the early detection of the virus will be a great way to stop this global pandemic from causing any more devastation and controlling its spread. South Korea implemented significant scale testing and decreased the number of cases per day to a very nominal value [4].

In this paper, we study and review the role of technologies like artificial intelligence and deep learning and their role in early detection and diagnosis of COVID-19.

2 Key Concepts

2.1 Data Science

Data science is one of the most sought after fields in the current software industry. There are many misconceptions about data science and its relationship with artificial intelligence models. Data science involves an assortment of several steps involved in the data science pipeline. These steps depend on the task assigned. So, how can we define data science? We can define data science as the science of collecting, storing, processing, describing, and modeling data. We can now discuss each of them and their implementation in the present scenario with proper examples.

Collecting data

Data collection and its role—Data collection is the process of collecting data(numerical, text, video, and audio) based on the question that is to be answered and the environment on which work is being done.

Environment type	1. Data is already collected	2. Data exists but is not owned and organized	3. Data needs to be collected
Technologies and skills involved	SQL queries, Java, Python	Web crawler, SQL	Programming, statistics

This step probably forms the background for all other upcoming actions. The collection of data from authentic and precise sources is necessary for designing effective models.

For example, in several models and designs for the detection of COVID-19, algorithms are made to differentiate CT scans and X-ray images of lungs of a COVID-19 patient with a healthy sample. In some cases, these are used for distinguishing between regular Pneumonia and COVID-19. These algorithms are nothing but neural networks that need to be trained on proper datasets. As of now, there are more than

three million cases of COVID-19 in the world, providing scientists with a large dataset to train the algorithms. The precision of detection depends on several factors. One of them is the number of parameters being considered during the evaluation. How are these parameters chosen? The parameters are determined such that all the internal and external symptoms can be covered, such as body temperature, the effects on blood sugar level, the oxygen level in the blood, and CT images of lungs. These can be considered as potential parameters. Most of the governments keep the data regarding these parameters in a structured manner and are available for research purposes. Most of these come under our category 1 and 2 in the above table. The available data can be accessed from the database by basic SQL queries and then manipulated according to use by languages like Java and Python. If our data falls under the second environment, we have to put some more effort and collect data from Web sites by Web scraping using Python. The third case is the most cumbersome, as in this case, data needs to be collected; in many instances, this data is collected by surveys and other traditional methods that prolong the process.

Storing data

In various projects related to the pandemic and its behavior, we mostly come across relational and operational databases.

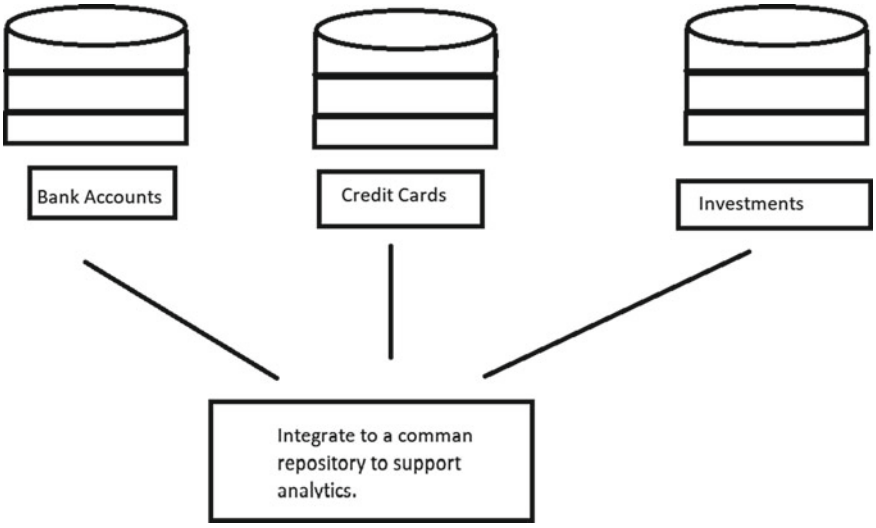
Patient ID	Name	Ailment	Contact details
00001	ABC	COVID	87955XXX14
00002	XYZ	COVID	1515XXXX25

The above table is an example of an operational or transactional database (the details are just for visualization purposes). In operational databases, data is arranged in the form of tables. Tableau and Excel are standard software used for the manipulation of operational databases. Generally, data such as patient records, medical examination reports, and drug dosage records are stored in operational databases. Operational databases are used to save data in real time.

Relational databases have predefined relationships between the data items stored. SQL is used for general queries and for gathering reports from data. Relational databases are helpful in many inferential statistics problems.

Data from multiple databases.

We are taking the example of a bank. A bank offers several services such as bank accounts, credit cards, and investments. If we want to study the preferences of customers in different services, we need to consider the data related to various services together.



Data warehouses consist of **structured** data from several databases integrated into a shared repository. Data lakes are similar structures except for the fact that data collected in them is raw and unstructured.

Processing data

Data wrangling or data munging

Above are the steps involved in data wrangling. In the first step, data is extracted from the database of the concerned organization. This data is mostly in tabular format. The information is then converted to some data interchange format, such as JavaScript Object Notation format (JSON). The data interchange format is then used to load data into the concerned algorithm.

Data cleaning

Data cleaning can be summarized in four primary steps:

- Filling missing values
- Standardizing keyword tags
- Correcting spelling and other manual errors
- Identify and remove outliners.

Data scaling, normalizing, and standardizing

Data scaling is done according to the project. Units for the quantities given in the procured dataset might not be suitable for the project calculations. We may scale the

data to suitable units. Normalizing is essential in the analysis of data. There might be cases where reporting mean and variance of the data using graphs might be difficult without normalizing it.

There are more than three million cases of COVID-19 in the world as of May 12, 2020. The dataset of such enormous size needs a lot of computing power. In these cases, performance becomes a critical issue. Distributed computing is quite useful in such cases. Softwares, such as Hadoop (Map-Reduce), expand can be used for Big Data analysis.

Describing Data

Describing data can be mainly divided into two parts:

- Visualizing data
- Summarizing data.

Tables and excel sheets are not suitable for visualizing data. Visualizing can be done using graphs and bar charts. For example, in later sections, we will be discussing several works related to the presentation of new data regarding the number of patients admitted, discharged, and deceased in real time. Excel sheets and tables are not feasible for representing data in such cases as it would be a cumbersome process to look for information in such format. Charts and graphs would be better as they provide a one look summary of the data to the common users.

When government agencies and research institutes are studying the situation of the pandemic, some common questions that pop up regularly are, "What is the typical number of cases registered daily?" "What is the typical variation in the number of cases registered daily?" This is where the summarizing of data steps in. Data can be summarized using specific quantities that define different queries regarding the data such as mean, median, mode, standard deviation, and variance. In the first question mentioned above mean of the cases registered can be the appropriate answer. Some of the important tools used here are descriptive statistics, iterative processes, and explanatory data analysis.

Modeling data

Statistical modeling-statistical modeling is used to identify underlying relationships between variables.

Let us suppose there is a new drug in the market; there are many questions associated with the medication that can be answered by statistics, like "Is the new drug effective in reducing blood sugar levels?" There are several techniques to answer these questions; one of them is by plotting these quantities against each other and finding the relationship between them. Suppose we get a normal distribution for the above quantities. Normal distribution can be defined using mean and variance only. In the case of COVID-19, a significant relationship is the one between age and number of days of treatment. Statistical modeling can be summarized in four points:

- Modeling underlying data distribution.
- Modeling underlying relations in data.

- Formulate and test hypotheses. (e.g., Is the drug effective?).
- Giving statistical guarantees.

Algorithmic modeling

It is an alternative to statistical modeling. Statistical modeling is used for sensitive fields like agriculture, where we have to give statistical guarantees. The drawback of statistical modeling is that we are left with only simple models, and we cannot take any complex relationships in the picture. Let y be the number of days required for the treatment of a patient; it is highly unlikely that the number of days of treatment will be related to only one parameter or that it will have a linear relationship with any of the parameters. Algorithmic modeling will help establish complex relationships between y and quantities such as blood pressure, body temperature, blood sugar level, age, gender, and other relevant parameters.

2.2 Machine Learning

The algorithms used in algorithmic modeling are machine learning algorithms. In machine learning, we choose very complex functions to show the relationship between the quantities. In the machine learning paradigm, real-world inputs and outputs are used to predict outputs for data that are new to the machine. Sometimes, the outputs may not be specified. There are several learning processes.

Supervised learning

The dataset consists of both inputs as well as outputs. We train the algorithm using raw input and target output. The trained algorithm is used to find the output for input data that is new to the algorithm.

Unsupervised learning

Unsupervised learning involves the training of dataset using raw input data without labeled outputs. Unsupervised learning looks for undetected patterns and correlations in the data without the help of any labeled outputs.

Reinforcement learning

Reinforcement learning is used to train algorithms using reward and punishment techniques. Raw input data is fed into the algorithm, and the system is rewarded for every correct output and punished for every wrong output. The trained algorithm is then used to detect output for new input.

Some important terms are used extensively in the machine learning paradigm:

Features—The measurable properties of a data object are called features. These are used as the input variable(s) for making necessary predictions. Features are essential for making efficient predictions about the data.

Example: Patient's age, blood sugar level, body temperature, etc.

Target/Label—The value that is to be predicted using machine learning is known as the target. It is the desired outcome of the machine learning exercise.

Example: Number of days required for recovery, chances of being diagnosed with COVID-19.

Model—The hypothesis that defines the relationship between the features and target is known as model.

Training-exposure of the model to features and expected targets. The process by which the machine establishes a relationship between the labels and the features.

Prediction-applying the model to unseen data and predicting the target (labels based on data).

Machine Learning Pipeline

Deep learning

Deep learning is part of machine learning. It consists of algorithms mainly based on artificial neural networks with representation learning. Deep learning is an essential asset for us when we have to deal with a large amount of high dimensional data. In this case, we use a specific class of complex ML models and algorithms, collectively known as deep learning.

2.3 Artificial Intelligence

We can study artificial intelligence and its role in any task under the following sub-sections:

Problem solving

In general problem-solving tasks, starting point and destination are provided, and we need to find a path linking the two. For example, tree algorithm can be used to solve a maze problem.

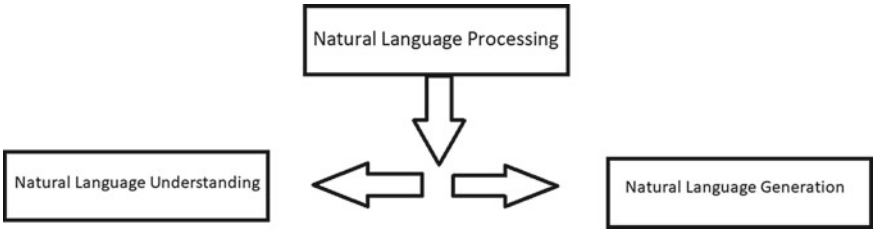
Knowledge representation

Knowledge representation is done using propositional and first-order logic (if-else statements).

Reasoning, decision making, and expert systems

Expert systems are used in various medical applications.

Expert systems have been employed in the identification of diseases for some time now. We can consider the example of dengue. A set of rules is specified based on symptoms of the disease. We define functions like hasRash(Patient), hasVomiting(Patient), hasHighFever(Patient), and target function hasDengue(Patient). All of these functions are generally of Boolean type. Now, how does it work? hasRash(Patient) AND hasVomiting(Patient) AND hasHighFever(Patient) → hasDengue(Patient). These functions have more logical functions working abstractly. For example, there might be an underlying function is TempGreater102(Patient) such that is TempGreater102(Patient) → hasHighFever(Patient). This function is also of the Boolean return type.



How can we pull this off? There are three significant steps in making expert systems.

1. Formulation of rules by domain experts.
2. Encoding the rules using knowledge representation.
3. Execution of rules and encoding done by a program.

Communication, perception, and actuation

Communication in AI is done by natural language processing; it consists of two sections. Natural language understanding is when a machine interprets human commands in the form of voice or handwritten text. Natural language generation is the generation of replies or responses by the device for the user. Perception is done by computer vision technology, microphones, and other audio-visual sources. Physical robots mostly do actuation. These robots are primarily data-driven hence involve a lot of concepts of data science. These robots can perform complex actuations by learning from simulations or by mimicking human examples.

2.4 Statistics

Statistics is the science of collecting, describing, and drawing inferences from the data. The population is the total collection of all objects that we are interested in studying. A sample is the subgroup of the population that we study to draw conclusions about the population. Generally, we are interested in some parameters of the sample.

How to select a sample?

"A sample and resulting statistic will be useful only if it is a correct and fair representation of the population." For example, let us assume a new medication for COVID-19 has come into the market, and we want to study its effect on the number of days required for treatment. If we take the sample as a group of university students here, then we will not get proper conclusions. The sample here is a group of university students; they have similar age, eating habits, sleep schedules, etc. Thus, the effect of the drug on them would not reflect the general effectiveness of the drug against COVID as it has to be used by a large number of people with different habits and lifestyle inclinations.

3 Contributions of AI and Data Science Against COVID-19 (Actual and Potential)

3.1 Visualization of the Pandemic

Data dashboards do visualization of the pandemic. Several data dashboards have been developed across the world to provide the world with the latest data about the number of people affected by the pandemic. The data provided by all of these dashboards is almost the same. The difference is the underlying machinery and the type of presentation. Presentation skills matter in case of visualizing a pandemic. These data dashboards are meant to be a resource for the public to analyze what level of precautions are to be taken. They must be highly readable, i.e., minimum effort is needed by the public for interpreting the data. These dashboards are nothing but landing pages with maps and visuals depicting where the virus has spread and the numbers on the latest recoveries and deaths. All databases are not the same; some are easy to navigate, some not all people look to same dashboards, and data is not available for all MIT technology review even posted a rank list of these dashboards. Many people raised their concerns regarding the safety of data of the patients, as the data of most dashboards is open-sourced. MIT technology review even posted a rank list of these dashboards.

Some of the best dashboards are:

Upcode

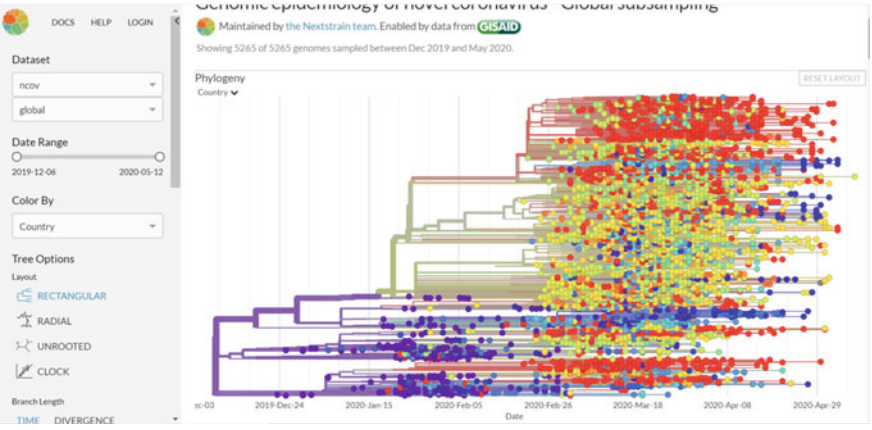
Upcode was ranked the best dashboard by MIT technology review. Upcode is not the best in terms of designs and looks, but it is very informative and easy to use. It uses the data provided by the Singapore Ministry of Health. It is quite transparent about data. The information from the dataset is compiled in the form of charts and graphs. The trends across gender, age, nationality, and location in the city are illustrated in a very understandable manner. The platform also shows the average days required for recovery.

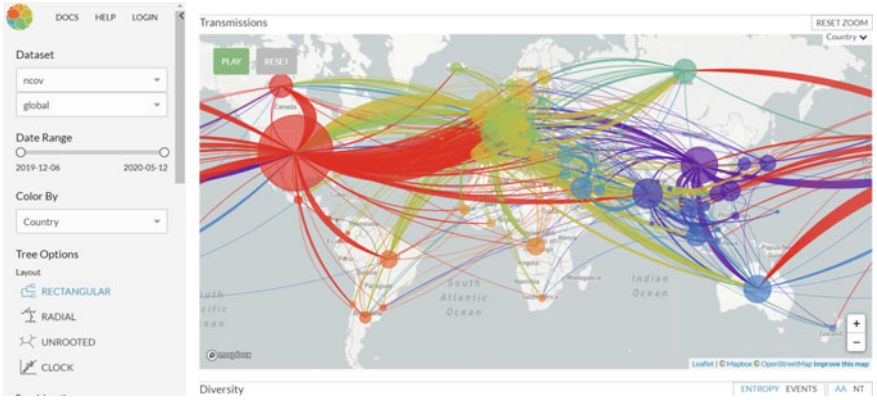
The above picture shows the number of cases with respect to time in the form of a bar graph. The pie chart shows active, critical, and recovered cases. Upcode maintains its data with utmost accuracy after rigorous verification.

There is only one problem with Upcode; it represents only Singapore, due to privacy reasons [5].

NextStrain

Trevor Bedford et al. have made a bioinformatics based dashboard. It is a bit more technical than most of the dashboards. The general public might have a little difficulty in interpreting it. However, researchers, scientists, and other enthusiasts might find it very useful. The team of epidemiologists and engineers at NextStrain collect data from the laboratories around the world, working on the genome of the virus, and integrate them into a central genome tree.

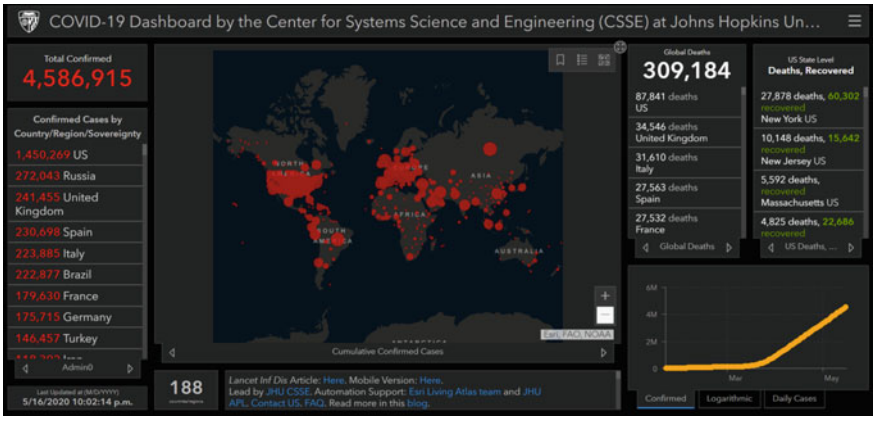




NextStrain provides a very impressive map depicting the status of transmission around the world. This map is one of its kinds. It gives governments and other higher authorities to control international travel from concerned countries, which could be very useful in controlling the infection [6].

JHU CSSE

The dashboard designed by JHU CSSE is inspired by the previous dashboard made for tracking measles in the USA. It is one of the dashboards with the highest utility value for people from different parts of the world. It provides precise information about the cases from almost every remote location on earth. Many other dashboards have taken inspiration from it.

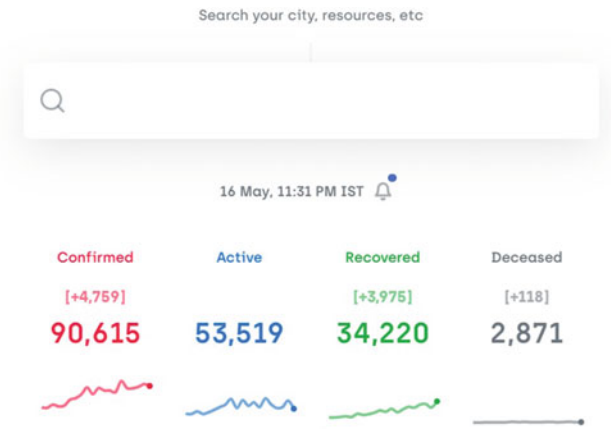


A screenshot of Dashboard designed by JHU CSSE

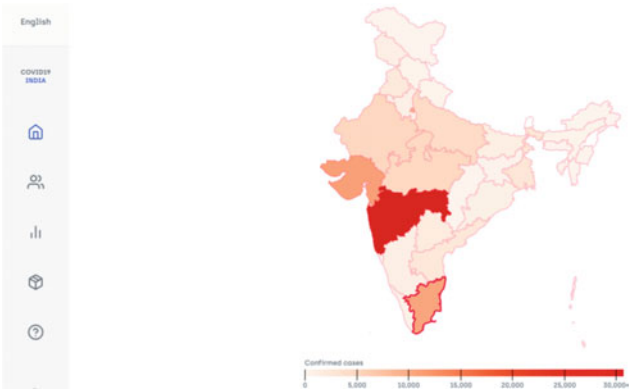
However, this dashboard has been updated thrice since its inception and can be considered as a work in progress [7].

3.2 India's COVID-19 Tracker

Similar to Upcode, a country centric dashboard was designed for India. India's COVID-19 tracker is different from other dashboards as it is updated at a faster pace than most of the dashboards in the world. This is because unlike other dashboards of its kind COVID-19 tracker is crowd-sourced. It gathers its information from state bulletins, press information bureau, and ANI reports, which are more recent as the Ministry of Health Affairs, India, updates its data at a scheduled time. It also has a section that suggests the nearest stores for buying essentials based on the selected location [8].



A screenshot of COVID 19 tracker's landing page



COVID 19 Tracker's Map indicating affected areas in India and severity using different color zones.

3.3 What Do All These Have in Common?

All the dashboards share several standard features such as the bar showing all the statistics about the cases registered, deceased, and recovered, the world map with the distribution of cases around the world, a chart showing the number of cases registered with time, etc. These are some essential pieces of information that are necessary for everyday people. Apart from these, some unique features like information about average recovery time, location of shops for essentials may be added.

3.4 Major Technologies Used

Data dashboards are mainly based on several data science concepts. Most of these concepts are discussed earlier in detail. We can relate most of them to the uses here. The data is collected mainly from sources of health ministries of different countries. Some dashboards, as we studied, were crowd-based (collecting data from press reports and other press resources) and some even scrape data from Web sites. The tools required are different for different sources. We can classify the sources mainly into three categories here:

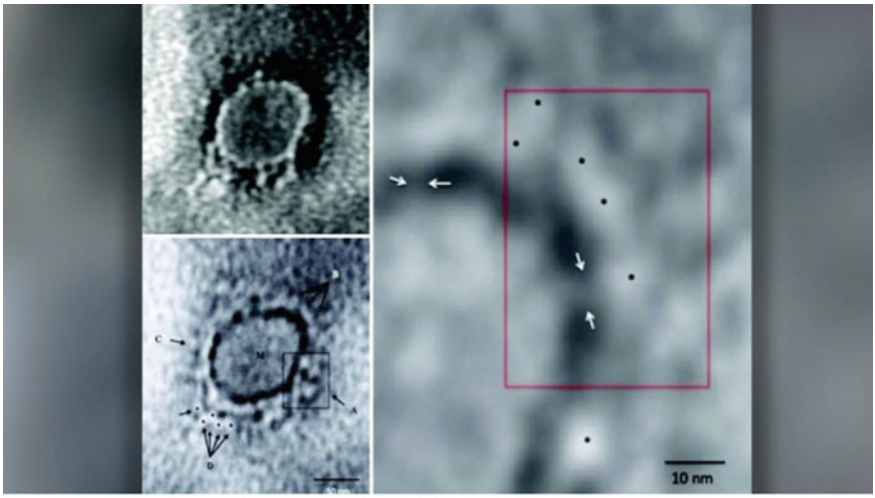
Environment type	1. Data is already collected Ex-data from databases of health ministries	2. Data exists but is not owned and organized Ex-data from other Web sites and dashboards	3. Data needs to be collected Ex-data from crowd-based sources such as press reports
Technologies and skills involved	SQL Queries, Java, Python	Web crawler, SQL	Programming, Statistics

Tableau, Python, or Excel can easily do the visualization part.

4 Early Detection and Diagnosis

Early detection and diagnosis of COVID-19 can be beneficial in controlling the spread of the disease. The mortality rate of COVID-19 low compared to other diseases like Ebola, HIV, etc. What is it about COVID that makes it more formidable? The problem with COVID-19 is that it has a long incubation period. It is highly contagious during this period, and the patient may not even know that he is suffering from COVID and might spread the infection to hundreds of people daily. Early detection and diagnosis will help significantly in controlling the disease. Early detection is also necessary so that treatment can be started as early as possible. The replication mechanism of the SARS-COV 2 is enabling it to mutate. Mutations make it difficult for the immune

system of the body to detect and fight against it. This is making it more difficult for scientists to come up with a vaccine for COVID-19.



Transmission electron microscopy imaging of Covid-19 | Photo from Indian Journal of Medical Research

The image shows the TEM image of coronavirus. The Indian Journal of Medical Research released it [9].

Artificial intelligence has been used earlier for the diagnosis of diseases like dengue. Image-based medical diagnosis will be beneficial in the identification of the illness, x-rays, and computed tomography can be used as inputs. The dataset for identification has grown significantly, with more than three million cases of COVID-19 around the world.

Reverse transcription-polymerase chain reaction (RT-PCR) is the most common method for diagnosis on COVID-19 being used these days. It is very costly and time taking, and sometimes inaccurate. Several machine learning methods are being used by researchers to find efficient improvements for RT-PCR tests.

Metsky et al. have used machine learning algorithms along with clustered regularly interspaced short palindromic repeats (CRISPR) technology to develop the assay designs and experimental details of 67 different viruses, including SARS-COV2. They are confident that this technology will be helpful in the fast detection of the disease and will lighten the burden from the diagnosis industry. The ML algorithms used are very specific and will eliminate the probability of false-positive cases. The authors state that this process will be useful not only for SARS-COV2 but for an extensive range of genomes [10].

Lack of accuracy in the testing facility is one of the most prominent reasons for the rapid spread of COVID-19. For example, if a person suffering from COVID is found to be negative due to inaccuracy can go on spreading the disease to hundreds of people daily.

RopezLincon et al. have used a different approach to improve the RT-PCR by using a convolutional neural network (CNN). These are used to classify the nucleic acid-based on associativity with SARS-COV2. SARS-COV2 is very similar to viruses such as SARS-COV1, MERS-COV, etc. Hence, identifying gets difficult. Missing information due to noise in the signal contributes to the difficulty. The convolutional neural network used here generates the features of the virus, including the genome sequence. The authors use a 21-base pair convolution over the whole genome (as opposed to previous approaches which only examine sequences of fixed length) and visualize the network's convolution and max-pooling layers to understand which particular sequences help to identify SARS-COV2. The authors also deploy a classification model (ex-logistic regression) to distinguish the cases into hospitalized and asymptomatic cases. The authors were able to identify SARS-COV2 with 99% accuracy [11].

Image-based medical diagnosis

We have discussed earlier that RT-PCR has several demerits such as limited resources for running tests, specimen collection, and inaccuracies. High costs of testing kits prevent developing countries from conducting tests on a large scale. RT-PCR is found to have accuracy as low as 60%–70%. Diagnosis using RT-PCR requires excluding several negative tests, and equipment being in short supply makes it very exorbitant. Novel coronavirus causes a severe infection in the lower respiratory tract in 50%–75% of the cases, which can be detected by CT findings [12].

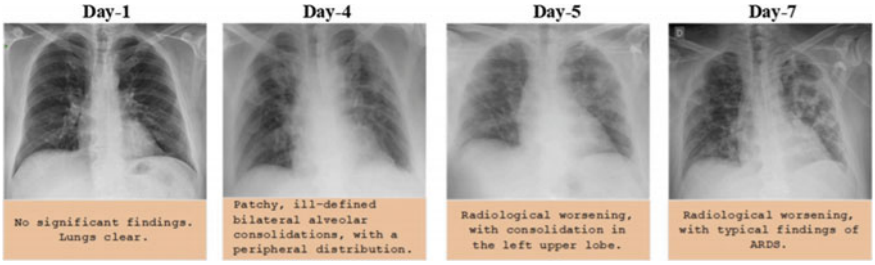
There are several radiological patterns that COVID-19 leaves; these can be found in several medical imageries. Identifying these patterns is difficult and time-consuming for even expert radiologists. ML algorithms can be used for fast analysis of the medical imagery, thus making it a strong candidate for fast diagnostic processes. Although the extent to which ML and medical imagery can be used for diagnosis is still under discussion [13, 14].

Machine learning brings several possibilities in the picture. Classification algorithms based on an artificial neural network can use deep learning that can be used to make binary (healthy vs. COVID) and multi-class (healthy vs. COVID vs. other types of Pneumonia) classifications. These approaches can use different architectures, such as ResNet [15], UNet ++ [16], and inception [17].

Ophir Gozes et al. have used automated AI-based tools for the analysis of CT images for the identification of COVID-19. The authors claim to state that this technique can be used to distinguish COVID-19 patients with non-patients. The authors used 2D and 3D deep learning models and combined existing AI models with medical imagery such as CT to create a system that can identify COVID with 98.2% sensitivity and 92.2% specificity. They used multiple international datasets, including datasets from affected areas in China. The testing dataset includes 157 patients [18].

X-ray images of the chest can also be used for COVID-19 detection. Machinery for X-ray imaging is more accessible and portable than that of CT imaging. X-ray imaging is also more cost-efficient than CT imaging. It can be used along with similar architecture (ResNet, convolutional neural network, etc.) as the one used in CT imaging techniques.

Turing Ozturk et al. have coupled artificial intelligence with radiological imaging for accurate detection of COVID-19. They presented a model that can automatically detect COVID-19 once raw X-ray is presented. The model is designed to perform binary (COVID vs. healthy) and multi-level classification (COVID vs. healthy vs. Pneumonia) with 98.08% accuracy for binary classes and 87.02% accuracy for multi-class classes [19].



Chest X-ray images of a 50-year-old COVID-19 patient with Pneumonia over a week [20]

Diagnosis of COVID is a very important and sensitive task; minute errors might result in catastrophic results. Image-based medical diagnosis has come up with satisfactory results, but it has a long way to go before we can consider it as a replacement for clinical diagnostic machinery. The image-based medical diagnosis must comply with certain standards that are set by medical associations worldwide. The models should be tested on larger, more diverse datasets. We observe that most of the papers we reviewed were tested on very small datasets and did not comply with most of the regulations set by medical associations. Therefore, these diagnosis techniques should be properly evaluated before implementing them in the medical diagnosis field [21].

5 Early Warnings and Alerts

In the case of early warnings and alerts, the Canadian AI model 'BlueDot' has done a commendable job in predicting the outbreak early. BlueDot is a Canadian health monitoring company. According to certain articles [22] and their statements, BlueDot predicted the outbreak of coronavirus pandemic in China and the world on December 31, 2019 and sent a report to its customers, after nine days on January 9, 2020, WHO declared COVID-19 a pandemic.

BlueDot collects a large amount of data from several sources like official statements of health organizations, press reports, digital media reports, airplane ticket records, and even animal health data. They use big data to handle such a huge amount of data. They use natural language processing and machine learning algorithms to isolate important data [22].

The MIT-based AI HealthMap used AI to map the spread of COVID-19 across the world. It took data from sources similar to BlueDot except for the fact that it took data from social media Web sites as well. The authorities in HealthMap clarified later that all the data that was procured and used was public. One of the most important facts here is that HealthMap predicted the outbreak even before BlueDot on 30th December itself. These AI models like BlueDot and HealthMap are very important tools for giving early warnings, but without human input, they cannot predict the intensity of the threat correctly. Let us take HealthMap; for example, HealthMap rated the COVID-19 threat to be 3 out of 5, and now more than 3 million people are suffering from it [22]. Thirty minutes after HealthMap, a scientist from ProMed (International society for infectious diseases) raised the alarm about COVID and understood its potential to become an epidemic. At the same time, it took the automated systems of HealthMap days to understand the potential threat posed by this pneumonia-like disease [23]. It can be concluded that though these systems are very powerful and fast, identification of threat level, and intensity of a problem needs human expertise for correct prediction.

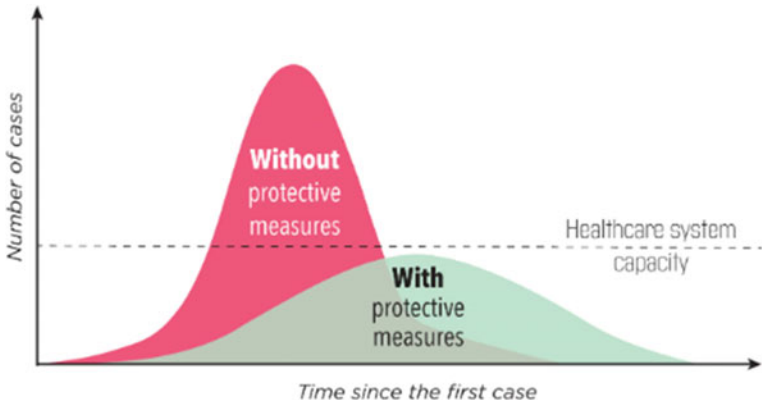
6 Tracking and Prediction

Artificial intelligence software can be used for tracking pandemics and their spread over different countries with time. In 2015, Zika virus had spread in Brazil, and eventually, in the rest of the Americas, this posed a major challenge to the health agencies there. They were not able to predict the spread of the virus and contain it later in 2019, and Mahmood Akhtar et al. have designed a dynamic neural network model to predict the outbreaks in real time. The model used the data from air travel details, socio-economic, and population data to predict the regions and states to be most affected by the virus. This model was applied to the Zika virus data, and it predicted the high-risk areas with an accuracy of 85% and a prediction range of 12 weeks [24]. Models like these can be used for tracking the COVID-19 pandemic; we need to train these ANN-based models with the new dataset of the current pandemic. Carnegie Mellon University has employed a similar model to predict the outbreak of seasonal flu; their team updates the model with a new dataset every year. This year they are updating the model with the dataset of COVID-19.

However, there are several challenges in implementing AI systems for tracking purposes in this situation. If we study the social media activity after the outbreak of COVID, it can be found that there is a lot of panic in the people regarding the sickness. Many AI systems worked by collecting the data from social media posts, the number of related searches on the Internet, etc. Due to the excess panic among people, searches related to symptoms are occurring unnecessarily even without any serious illnesses. Symptoms of COVID are contributing to this problem as some of the symptoms of COVID overlap with symptoms of several diseases that are not that serious [25].

Lack of historical data is another problem that poses a challenge to implement these technologies. The rate of spread of COVID-19 is way higher than any other pandemic that humanity has observed. The Spanish flu of 1918 had a similar contagious nature, but the records of that pandemic are not readily available, and the virus causing it was also quite different from SARS-COV2.

Tracking and prediction will help the governments in analyzing, planning, and preparing for the pandemic. It will also be helpful in finding out how much the preventive measures such as quarantine, lockdowns, and social distancing are working for flattening the epidemiological curve.



A curve showing the number of cases with time with and without the protective measures [26].

6.1 Clustering

The outbreak of COVID in any region or country depends on several factors like demographics, air travel history, government policies, climate, etc. On the basis of these factors, countries, and regions can be divided into clusters. Data from one of the regions can be used to predict the scenario of the outbreak in other countries.

Carrillo Larco et al. have used unsupervised machine learning algorithms (k-means) to define clusters for countries based on the level of air pollution, socioeconomic coverage, and healthcare infrastructure. The model classified 155 countries into clusters. The model worked well in classifying the countries in terms of the number of cases but in terms of the number of deaths or fatality rates [27].

Zixin Hu et al. have studied data from 31 provinces in China fed to a modified autoencoder (MAE) for broadcasting purposes. The authors then extracted the data from the autoencoder's latent data layers and fed it into a k-means clustering algorithm, which grouped them into clusters for further analysis.

7 Patient Outcome Prediction

The number of patients suffering from COVID-19 is increasing day by day; we have seen the top-rated healthcare systems such as that of Italy collapse because of the overwhelming pressure of the ever-increasing number of cases. We observe that most of the cases consist of mild symptoms and may even go undetected, but some of the cases may develop into acute respiratory distress syndrome (ARDS). AI can help in analyzing and narrowing down factors that lead to ARDS and, eventually, death due to respiratory failure. There have been several papers suggesting the use of medical data such as blood tests, age, and other medical data for evaluating the risk of development of ARDS at the later stages. Several clinical indicators have been identified by ML methods.

Li Yan et al. have studied electronic records of more than two thousand patients and screened out records of twenty-nine patients. They tested these records using a prediction model based on the XGBoost machine learning algorithm. Their model is shortlisted lactic dehydrogenase (LDH), lymphocyte, and high-sensitivity C-reactive protein (hs-CRP) out of more than three hundred features. The authors claim that their model predicts mortality of a patient with more than 90% accuracy [28].

Several studies show that medical imagery, like X-ray and CT imaging, can be used for patient outcome prediction. Zhenyu Tang et al. have used CT imaging for severity assessment. The number of cases is increasing rapidly, and manual severity assessment will become extremely difficult and will eat up an important time of treatment. The authors claim that their model to have an accuracy of 0.875 [29].

8 Fake News and Misinformation

According to WHO, an infodemic may be defined as a scenario where there are surplus sources of information with unknown accuracy, which may be problematic for people who are looking for correct information when they need it. Social media can be our most important tool against this pandemic if we use it correctly. The promotion of social distancing and other preventive measures using social media can be very useful. The social media platforms host accounts of almost two billion active users. These numbers give us an estimate of their effect. There are several posts on social media propagating misinformation, selling fake coronavirus cures; there have even been several cyber attacks on databases of hospitals. The United Nations has warned people to verify any information regarding vaccines and cures with authentic sources such as the Web site of WHO [30].

Artificial intelligence can be used for the detection of fake news regarding COVID-19. Adrian Groza used description logics (DLs) to differentiate between the news on the basis of the source of information. The system detects inconsistencies between trusted and non-trusted medical sources [31].

9 Treatment and Cure

AI is being used in the pharmaceutical and drug-making industry for a long time before the outbreak of COVID-19. AI can be used for easing several tasks such as studying the structure of different viruses and proteins, studying the effects of different drugs on viruses, etc. With these extremely helpful features, AI can help in resurfacing past drugs that may be effective in this scenario or even discovering new drugs that may be effective against this disease. Google's AI software DeepMind predicted the structure of proteins associated with SARS-COV2. However, they specify that there is no experimental verification about the accuracy of the predicted structures. The structure of proteins is a very important asset for making a useful drug. The Web site also mentions that the system correctly predicted the structure of the protein spikes on the SARS-COV2, which was verified using the data from the Protein Data Bank, which signifies that the accuracy of the software is better compared to any previous software [32].

Bo Ram Beck et al. have used deep learning for drug resurfacing for discovering effective treatment of COVID-19. The authors used a deep learning-based drug target interaction model called the molecular transformer-drug target interaction model (MT-DTI). The model showed that atazanavir, an antiretroviral that was used for preventing human immunodeficiency virus (HIV), is one of the best candidates against SARS-COV2. Molecule transformer-drug target interaction (MT-DTI) predicted the effectiveness of the drug on the basis of affinity between the drug and the protein spikes of the virus. Scientists at benevolent AI, an AI company from the UK identified Baricitinib, a drug used for rheumatoid arthritis could be used for the treatment of COVID-19 [33].

Konstantin Avchaciov et al. have used deep learning methods to find a number of drugs that could be used against COVID-19, including some of the drugs currently used for lung cancer treatments [34].

The drugs that we have discussed are not very likely to be used in the near future. The main reason behind this being a large number of checks and screenings that drugs have to go through. Every drug has to satisfy certain regulations set by the medical associations of the world before coming into the market for general use. All these screening processes take a lot of time.

10 Resources for Data

Machine learning and deep learning models require a very large amount of data and computing power to create effectively functional algorithms.

Data from the cases

Almost every section that we have discussed in this paper requires data about the number and location of the cases. Several datasets for these purposes are provided

by organizations such as WHO and Centers for Disease control for every country. These have been hosted on public repositories by organizations like GITHUB and institutions like Johns Hopkins JSSE. Apart from location and number, there are several complementary datasets that may or may not be available in a structured form, such as the socio-economic impact of the virus and the perception of people about the virus. Efforts are being made to extract and study the de-identified large-scale data to study the local impact and evolution of the disease in Italy [35] and North America [36].

Data from Scientific Research

Machine learning methods can be used to extract and explicate data about coronavirus from written materials. The publications about transmission, incubation, and stability of the viruses, vaccines, and health care are available on several databases like WHO global research database on COVID-19. We also have the COVID-19 Open Research Dataset (CORD-19), which currently has over 52,000 relevant research articles making it the largest open dataset available. NLP techniques can be applied to develop text mining tools and resources that can help extract data for the medical community to find answers to key scientific questions regarding the nature and progress of COVID-19.

Data from social media

Several times the data from social media and news reports can be used to complete the data for scientific purposes. (Datasets like COVID-19 Twitter dataset) These datasets are maintained by tweets about COVID-19; these may be helpful in tracking down the spread of misinformation about the pandemic. The dataset COVID-19 real-world worry dataset is maintained with the emotional responses of people about the pandemic. These datasets are useful in judging the reaction of the people toward the pandemic, for preventing panic and unnecessary social gatherings.

There are repositories like the COVID-19 Coronavirus News Article database and the COVID-19 Television Coverage Dataset that can be used to study the working and evolution of the paper and television media as the pandemic advances.

Clinical data

Datasets for clinical data are not as readily available as other datasets. Some CT scans and X-ray images that we discussed in the medical imagery diagnosis section may be open-sourced and available to the public. Efforts are being made to make relevant data crowd-sourced and open-sourced such as the COVID chest X-ray dataset. However, using and maintaining these datasets are quite difficult. We have discussed several steps involved in the data science pipeline; in this case, the steps like data collection and model training may be performed by a computer scientist, but data cleaning, giving annotations and data labeling require professional medical expertise and may be performed only by doctors and clinicians. To fulfill the need for this much-needed data, the number of initiatives and repositories are increasing the Data4COVID Living Data Repository [37] and the COVID-19 Dataset Clearinghouse [38] some of them.

11 Conclusion

This review shows that ML and AI have the potential to contribute to a wide range of domains against the COVID-19 pandemic. Particularly, we have discussed their role in visualization, diagnosis, tracking and prediction, drug delivery, patient outcome detection, and detection of fake news. However, we note that most of the models do not have operational maturity at this stage. To improve this scenario, we need to take certain steps as a community. Firstly, more open-sourced repositories should be available with relevant datasets; secondly, as we have observed that we need interdisciplinary co-ordination to make proper models. Therefore, medical professionals and data scientists need to coordinate their work. Thirdly, all the sectorial and international differences should be set aside, and different companies, countries, and research institutes should join hands to achieve a common goal. We hope that this review will help the research community in knowing where we are standing know and how much more do we have to work. We also hope that this review will provide an insight into the areas that are achievable using data science methods and show us the domains where these methods might be nothing more than a great risk.

References

1. Situation Report 1 WHO https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200121-sitrep-1-2019-ncov.pdf?sfvrsn=20a99c10_4 Accessed 1st May 2020
2. Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., et al.: Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology **2020**, 200642 (2019)
3. Scroll.in Why does Coronavirus Test cost 4500 in India? https://scroll.in/article/961002/why-does-the-coronavirus-test-cost-rs-4500-in-india Accessed 14th May 2020
4. Trust, Testing and Tracing: How South Korea succeeded where the US stumbled in coronavirus response. https://abcnews.go.com/Health/trust-testing-tracing-south-korea-succeeded-us-stumbled/story?id=70433504 Accessed 14th May 2020
5. Dashboard for COVID 19 outbreak in Singapore. https://www.againstcovid19.com/singapore/dashboard Accessed 16th May 2020
6. Genomic epidemiology of novel coronavirus - Global subsampling https://nextstrain.org/ncov/global Accessed 16th May 2020
7. Dashboard by the Center for Systems Science and Engineering (CSSE) at John Hopkins University. https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6 Accessed 16th May 2020
8. Coronavirus outbreak in India https://www.covid19india.org/ Accessed 16th May
9. Prasad, S., Potdar, V., Cherian, S., Abraham, P., Basu, A.: ICMR-NIV NIC Team. Transmission electron spectroscopy of SARS-COV-2 Indian J. Med. Res. **151**(2 & 3), 241–243 (Feb and Mar 2020)
10. Metsky, H.C., Freije, C.A., Kosoko-Thoroddsen, T.S.F., Sabeti, P.C., Myhrvold, C.: CRISPR-based surveillance for COVID-19 using genomically-comprehensive machine learning design. bioRxiv preprint bioRxiv:20200226967026. 2020
11. Lopez-Rincon, A., Tonda, A., Mendoza-Maldonado, L., Claassen, E., Garssen, J., Kraneveld, A.D.:Accurate identification of Sars-Cov-2 from viral genome sequences using deep learning. bioRxiv preprint bioRxiv:20200313990242v1. 2020

12. Kanne, J.P., Little, B.P., Chung, J.H., Elicker, B.M., Ketai, L.H.: Essentials for radiologists on COVID-19: an update—Radiology Scientific Expert Panel. Radiology. p. 200527 (2020)
13. Ng, M.Y., Lee, E.Y., Yang, J., Yang, F., Li, X., Wang, H. et al.: Imaging profile of the COVID-19 infection: radiologic findings and literature review. Radiology: Cardiothoracic Imaging. **2**(1), e200034 (2020)
14. Weinstock, M.B, RJeaEchenique, A.: Chest x-ray findings in 636 ambulatory patients with COVID-19 presenting to an urgent care center: a normal chest x-ray is no guarantee. J. Urgent Care Med. p. 13–18 (2020)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, p. 770–778 (2016)
16. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D. et al.: Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
17. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet ++: A nested u-net architecture for medical image segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, p. 3–11. Springer, 2018
18. Gozes, O., Frid-Adar, M., Greenspan, H., Browning, P.D., Zhang, H., Ji, W. et al.: 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: 200305037. 2020
19. Ozturk, Tulin, Talo, Muhammed, Yildirim, Eylul Azra, Baloglu, Ulas Baran, Ozal Yildirim, U., Acharya, Rajendra: Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput. Biol. Med. **121**, 103792 (2020). https://doi.org/10.1016/j.compbiomed.2020.103792. ISSN 0010–4825
20. Edgar Lorente, COVID-19 pneumonia—evolution over a week. https://radiopaedia.org/cases/COVID-19-pneumonia-evolution-over-a-week-1?lang¼us. Accessed 16th May
21. Nagendran, M., Chen, Y., Lovejoy, C.A., Gordon, A.C., Komorowski, M., Harvey, H. et al.: Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. bmj. 368 (2020)
22. ARS Electronica https://ars.electronica.art/index.html Accessed 16th May 2020
23. Can AI flag disease outbreaks faster than humans? Not quite AP News https://apnews.com/100fbb228c958f98d4c755b133112582 Accessed 16th May 2020
24. Akhtar, M., Kraemer, M.U.G., Gardner, L.M.: A dynamic neural network model for predicting risk of Zika in real time. BMC Med. **17**, 171 (2019)
25. Artificial Intelligence against COVID-19: an Early Reviewhttps://towardsdatascience.com/artificial-intelligence-against-covid-19-an-early-review-92a8360edaba Accessed 16th May 2020
26. Relief Central Epidemic(Epi) Curves for Coronavirus COVID 19 https://relief.unboundmedicine.com/relief/view/Coronavirus-Guidelines/2355041/all/Epidemic__Epi__Curves_for_Coronavirus_COVID_19 Accessed 16th May 2020
27. carrillo-larco r, castillo-cara m. using country-level variables to classify countries according to the number of confirmed covid-19 cases: an unsupervised machine learning approach [version 1; peer review: awaiting peer review]. welcome open research. 2020;5(56). https://doi.org/10.12688/wellcomeopenres.15819.1
28. Zixin Hu, Qiyang Ge, Shudi Li, Li Jin, MomiaoXiong, Artificial Intelligence Forecasting of Covid-19 in China. arXiv, 2020
29. Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan Li Yan, Hai,Tao Zhang, Yang Xiao, Maolin Wang, Chuan Sun, Jing Liang, Shusheng Li, Mingyang Zhang, Yuqi Guo, Ying Xiao, Xiuchuan Tang, Haosen Cao, Xi Tan, Niannian Huang, Bo Jiao, Ailin Luo, Zhiguo Cao, Hui Xu, Ye Yuanmed Rxiv 2020.02.27.20028027; doi:https://doi.org/10.1101/2020.02.27.20028027
30. Tang, Z., Zhao, W., Xie, X., Zhong, Z., Shi, F., Liu, J. et al.: Severity assessment of coronavirus disease 2019 (COVID-19) using quantitative features from chest CT images. arXiv preprint arXiv:200311988. 2020

31. WHO website Emergencies Coronavirus disease https://www.who.int/emergencies/diseases/novel-coronavirus-2019?gclid=CjwKCAjw_LL2BRAkEiwAv2Y3SbUEvAwKpNUpsJPSreZUGtQGlCBedIXdmXmOirYAiNuXCjbSR5VAtxoCqjoQAvD_BwE Accessed 17th May 2020
32. Adrian Groza, Detecting the fake news regarding Coronavirus by reasoning on COVID 19 ontology arXiv:2004.12330 2020
33. DeepMind https://deepmind.com/research Accessed 17th May 2020
34. Beck, B.R., Shin, B., Choi, Y., Park, S., Kang, K.: Predicting commercially available antiviral drugs that may act on the novel coronavirus (2019-nCoV), Wuhan, China through a drug-target interaction deep learning model. bioRxiv preprint bioRxiv:20200131929547. 2020
35. Potential new treatment for COVID 19 uncovered by BenevolentAI enters trials https://techcrunch.com/2020/04/14/potential-new-treatment-for-covid-19-uncovered-by-benevolentai-enters-trials/ Accessed 17th May 2020
36. COVID 19 Mobility Monitoring Project https://covid19mm.github.io/in-progress/2020/03/13/first-report-assessment.html Accessed 17th May 2020
37. Safegraph US consumer activity during COVID 19 Pandemic https://www.safegraph.com/dashboard/covid19-commerce-patterns?is=5e7a3815f20d617a17a33173 Accessed 17th May 2020
38. DataforCOVID Repository https://docs.google.com/document/d/1JWeD1AaIGKMPry_EN8GjIqwX4J4KLQIAqP09exZ-ENI/edit Accessed 17th May 2020