



Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Exploratory study of existing approaches for analyzing epidemics

Neetu Sardana, Arpita Jadhav Bhatt

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA, INDIA

Key points the chapter describes

1. Describes openly available pandemic datasets and their characteristics.
2. Discusses an important metric, the reproductive number, which can help in finding the best estimation of parameters that impact the prediction maximally.
3. Highlights the importance of Epidemic models for studying infectious diseases and describes the commonly used prime models.
4. Machine learning (ML) models are the popular choice for studying the epidemic model. This chapter portrays the epidemic dissemination factors used by these models in health predictive analysis.
5. Provides varied visualization charts that can help in understanding the patterns of virus spread in a pandemic situation.
6. Describes the semantic techniques used to gain new insights and find answers for complex research queries related to pandemics

4.1 Introduction

The recent novel coronavirus pandemic has created panic across the world and impacted the lives of several people. It has forced governments to impose lockdowns, travel bans, and large-scale quarantines as a preventive measure to reduce its impact. It has been declared a pandemic by the World Health Organization (WHO). A pandemic is an infectious disease that spreads across countries. Often the term “pandemic” is replaced by the term “epidemic,” which represents the spread of an infectious disease within a wide geographical area. An epidemic is generally a short-term outbreak of a pathogen. The spread of an infectious disease consists of disease-related features such as infectious agents; mode of transmission; latent period; infection time period; and social, demographic, economic, and geographical factors associated with the virus. Various epidemic diseases have been reported in the past. To substantiate the fact, the Zika outbreak was reported in the year 2016. It had raised an alarm in

the United States as it trapped 218 individuals in Florida and 6 individuals in Texas. Even the disease HIV had spread across the planet in the past and created a panic situation. Influenza is becoming more contagious every year. Health officials commonly refer to seasonal outbreaks as pandemics. The novel pandemic Influenza H1N1 2009 occurred in the United States and affected approximately 60 million people, causing 274,304 hospitalizations and 12,469 deaths (Chan, 2010). Another epidemic, the cholera outbreak, initiated in India and reached across continents, and over one million lives got affected across the world. Table 4–1 lists the major pandemics that have occurred from the year 165 until 2020 (LePan, 2021).

Despite the persistence of pandemics since the year 165, it can be observed from Table 4–1 that there is a consistent gradual reduction in the death rates. Healthcare professionals strive hard to understand the factors that incubate pandemics to mitigate their impact. In understanding and mitigating the viruses' impact, mathematical and statistical models can help. They can also help in understanding the risks associated with an infectious disease. Researchers and scientists model the behavior of viruses using a wide range of techniques such as ML, artificial intelligence (AI), or epidemic-based techniques.

In modeling any epidemic, it is extremely important to analyze and predict its impact. Researchers and scientists have conducted a lot of studies on analyzing the pattern of virus spread and estimate its future course. These studies have helped in modeling the behavior of viruses in terms of their transmission and distribution which describes the pattern and speed of virus spread. Besides, modeling of the virus helps in planning the control measures and vaccination policies and observing its future impact. Car et al. modeled the COVID-19 virus and used infected, recovered, and dead patients from 406 places to demonstrate that AI Models can be used to model the spread and effect of an infective disease even with a limited dataset, which means that to achieve encouraging results, it is not mandatory for AI algorithms to consider long periods of data (Car et al., 2020). Pratchaya et al. performed

Table 4–1 List of major pandemics.

Pandemics	Year
Plague	1665
Cholera	1817–1923
Yellow Fever	Late 1885
Russian Flu	1889–1890
Spanish Flu	1918–1919
Asian Flu	1957–1958
Hong Kong Flu	1968–1970
HIV AIDS	1981–Present
H1N1 (Swine Flu)	2009–2010
SARS	2002–2003
Ebola	2014–16
MERS	2015
Covid-19	2019–Present

modeling for the Dengue virus using the susceptible-infected-recovered (SIR) technique (Chanprasopchai et al., 2018). The study was performed to determine the role of dengue vaccination. It has been found that hospitalization time can be reduced if an infected person is vaccinated in comparison to those who are not vaccinated. Researchers have studied the pandemic using visual graphs. Ross Maciejewski et al. developed a visual analytics toolkit (PanViz) to model, analyze, and explore the influenza pandemic using population size, demographics, and the number of hospital beds (Maciejewski et al., 2011). PanViz assisted in computing the population dynamics and infection rate. It enabled exploring the impact of the pandemic on the population through the geographical spatiotemporal view. According to the WHO, a pandemic initiates when three conditions are met: first, a new virus originates; second, it infects individuals and causes severe sickness; and third, it breaks out easily and sustains among individuals. This chapter gives a detailed overview of the existing approaches as well as tools to analyze varied epidemic situations. The chapter is organized in the following manner: Section 4.2 describes the available dataset repositories on pandemics, Section 4.3 describes the available dataset repositories on pandemics, Section 4.4 highlights the various visualization techniques available to analyze data on pandemics, Section 4.5 details the epidemic modeling techniques for the analysis and prediction of pandemics, Section 4.6 describes machine learning approaches for the analysis and prediction of pandemics, Section 4.7 explains the semantic technologies, and Sections 4.8, 4.9, 4.10 present summary, review questions, and discussion questions respectively. The chapter ends with a problem statement. The chapter ends with a summary.

4.2 Pandemics dataset repositories

This section describes a few popularly used pandemic datasets for varied research-related purposes. The datasets can be used for predictive modeling. They can also be used to decipher patterns.

4.2.1 COVID-19

The novel COVID-19 pandemic has changed our lives entirely. It has caused unprecedented public health and economic crises. The Johns Hopkins Coronavirus Resource Centre (CRC) is a rich source of the COVID-19 dataset which is continuously updated (Johns Hopkins University, 2020). The websites of John Hopkins University, the WHO, and Ding Xiang Yuan are authorized by the Chinese government. The sites report the number of confirmed, recovered, and dead COVID-19 cases in the affected countries and regions. The repository of Johns Hopkins University Centre for Systems Science and Engineering (JHU CSSE) is available as a GitHub repository (Johns Hopkins University, 2021). This dataset has been popularly used among researchers for predictive modeling and analysis.

4.2.2 Zika virus

Zika virus is a mosquito-borne virus. It is caused by the bite of infected *Aedes*, which is a type of mosquito species. It remains active throughout the day and can infect a person in the daytime. Zika pandemics occurred in 2016. The Zika virus dataset has been released by the Centers for Disease Control and Prevention (CDC) on Kaggle for exploration ([Kaggle, 2021d](#)). This Kaggle dataset has details on, for instance, when the case was clinically confirmed, lab-confirmed, suspected, etc. The dataset is available state wise and country wise. It is in Excel format. The Zika virus datasets, which are immensely used, have also been made available by World Data Atlas ([World Data Atlas](#)). It consists of Zika cases in all countries and territories with active Zika virus transmission.

4.2.3 H1N1 influenza

The H1N1 is an influenza pandemic that occurred in 2009 and lasted until 2010. It was detected in the United States and later spread across the world. This new H1N1 virus contained an exclusive mixture of influenza genes not identified in animals or people in the past. This virus was known as influenza, an H1N1pdm09 virus. According to the CDC, there were 60.8 million cases including hospitalizations and deaths. The dataset for the influenza virus is available at Kaggle, which has been reported by the WHO ([Kaggle, 2021c](#)). Twenty sets of influenza datasets are available in data world ([Data World, 2020](#)).

4.2.4 Ebola

Ebola is a deadly virus that causes fever, body aches, diarrhea, and even damages to the immune system and organs. The Ebola virus affected several African countries from the year 2013 until 2016. The virus also spread to neighboring countries. It caused a significant number of deaths, with a fatality rate of 40% and a total of 28,616 cases and 11,310 deaths. The Ebolavirus outbreak data are available on Kaggle ([Kaggle, 2021a](#)). The file contains confirmed cases and deaths from August 2014 to March 2016 from across the world. The records in the file represent the region/location for each day.

4.2.5 Acquired immunodeficiency syndrome

Acquired immunodeficiency syndrome (AIDS) is a life-threatening condition triggered by the human immunodeficiency virus (HIV). HIV damages the immune system. During the time of epidemics, the status of AIDS across the world can be analyzed from the data available in the Kaggle repository ([Kaggle, 2021b](#)). The data are provided by the WHO and UNESCO. It contains the details related to the number of people with AIDS, total deaths due to HIV, and prevention of mother-to-child transmission estimates. Especially, the data can help in analyzing where each country stands corresponding to HIV/AIDS. The European Union Open Data Portal (EU ODP) gives access to AIDS datasets for commercial and noncommercial purposes. They have also provided the HIV infection and AIDS data for research purposes ([Union, 2021](#)).

4.3 Analysis of pandemics using visualization

This section describes the various charts or graphs such as line charts, bar charts, histograms, and pie charts that can help in analyzing and understanding the social well-being as well as economic consequences of pandemics. The human brain is capable of processing images faster when compared to processing text. It processes pictorial information much faster than written text. Therefore the graphical representation of data plays a very important role in interpreting pandemic situations.

As the outburst of the novel coronavirus has progressively spread across the globe, epidemiological data need to be understood in such a manner that the model of statistical data analysis as well as visualization increases the understanding of the pandemic situation among people. Through visualization, one can keep track of the possible epidemic hotspots that might require more attention from the authorities. Data visualization techniques can be used to compare the effectiveness of different social measures taken by different countries to contain the distribution of pathogens. For example, it could be used to compare the effectiveness of centrally imposed lockdown in different countries and later study the effects of mitigation measures. One can perform numerous analyses from the data provided during an epidemic such as testing rates, infection-to-death ratio, and age-wise vulnerability to answer many uncertain questions related to an epidemic.

Data visualization plays a vital role in convincing people to change their behavior. Once a virus spreads across several areas, public health officials need to make important decisions about how much information they need to communicate to the public and at what point. One of the utmost significant aspects of containing an outbreak is convincing people to adapt their behavior when it is not proximately clear what they are supposed to do in case of a pandemic situation. In recent pandemic situations like COVID-19, data visualization has played an important role in communicating as well as convincing people. We have all seen models predicting and analyzing the spread of the novel coronavirus, most notably the “flatten the curve” graphic that has been widely shared online and on television and shared across social media. Data representations using charts, bar graphs, and line graphs are a better way of representing data rather than representing data using compound and overwhelming datasets that are mined during a pandemic. When users are dealing with quantitative data or datasets that are so large, (with the number of cases in the millions), the data visualization technique helps people to relate in such a manner that they can understand the things happening around them. The technique is one of the fastest as well as the most efficient ways to wrap your mind around huge amounts of data. Data visualization becomes a medium to create awareness among common people.

Visualization techniques can be categorized into two groups: (1) visualization with nil or less interactivity and (2) complex and interactive visualization techniques using tools (Biswas *et al.*, 2020). In the past, several static visualization techniques have been employed, which used bar graphs and line graphs to represent different attributes of the COVID-19 dataset. The Washington Post has utilized a collection of bar graphs to compare new active cases and bereavements between different countries. In their work, every chart denoted a country and

respectively each bar of every chart displayed the number of COVID-19 cases or deaths (Biswas et al., 2020). Kylie Stevens combined bar and line charts to construct a composite static graph to represent two diverse data attributes. They used a bar chart for displaying the number of new cases and the line charts were used to represent the percentage change of deaths encountered daily. They annotated their charts to depict several events such as the closing of schools and complete lockdown. The Financial Times also developed a data visualization technique with a sequence of line charts to compare daily deaths in a region. Additionally, a majority of websites have utilized interactive world map diagrams or the world globe as the main layout for visualizing the COVID-19 cases (Biswas et al., 2020). The visualizations had mouse hover as well as click interactivity to display several attributes, namely the number of active cases, deaths, recovered cases, total cases reported, fatal cases, etc. Flourish studio employed four different types of data visualization techniques to manage distinct information to update live graphics. Interactive stacked bar charts were employed by Wikipedia and Tableau Public to depict attributes such as the counts of total cases, deaths, recoveries, active cases, and hospitalized people. The International SOS website employed an interactive trajectory chart to compare the rates of new cases daily and determine the outbreak of the virus in several countries. The WHO and websites hosted by Bing used a novel approach to link the total number of COVID-19 cases with the radius of a semitransparent circle.

Mike et al. have visualized the pandemic risk using network analysis (So et al., 2020). The authors have provided a direct visualization of the pandemic risk informative figures using network-connected graphs where several regions are denoted as nodes. Based on the number of reported confirmed cases of COVID-19 in these regions, they have analyzed the degree of connectedness among them to estimate the pandemic risk. They have constructed a pandemic network to study the relationship of changes based on the number of confirmed cases. Network analysis has been applied to understand the dynamics behind the spread of this infectious virus. During their analysis, the authors retrieved several COVID-19 cases that were confirmed in China for the period of late January 2020 to March 2020. Later plots of the established cases were created for nine time periods. Fig. 4–1 depicts plots of the confirmed COVID-19 cases which were constructed by Mike et al. for different periods using bar graphs and network graphs (So et al., 2020).

In addition to the above, the authors also collected the number of confirmed cases worldwide including in China for the same period, and plots were constructed. Network graphs were built depending on the correlation of changes concerning the number of confirmed cases amid two geographical areas. Fig. 4–2 depicts a snapshot of daily COVID-19 cases reported worldwide and network graphs (So et al., 2020).

The authors identified that two geographical areas are connected in the network if the correlation is greater than 0.5. The pandemic risk was visualized based on how regularly different areas were linked in the network graph. The study analyzed if the regions are more frequently connected, the density of the lines in the graphs increases, and thus the propensity of the coevaluation of virus spread among these highly correlated regions increases. The authors have used different network graphs to determine the increase or decrease in the pandemic risk

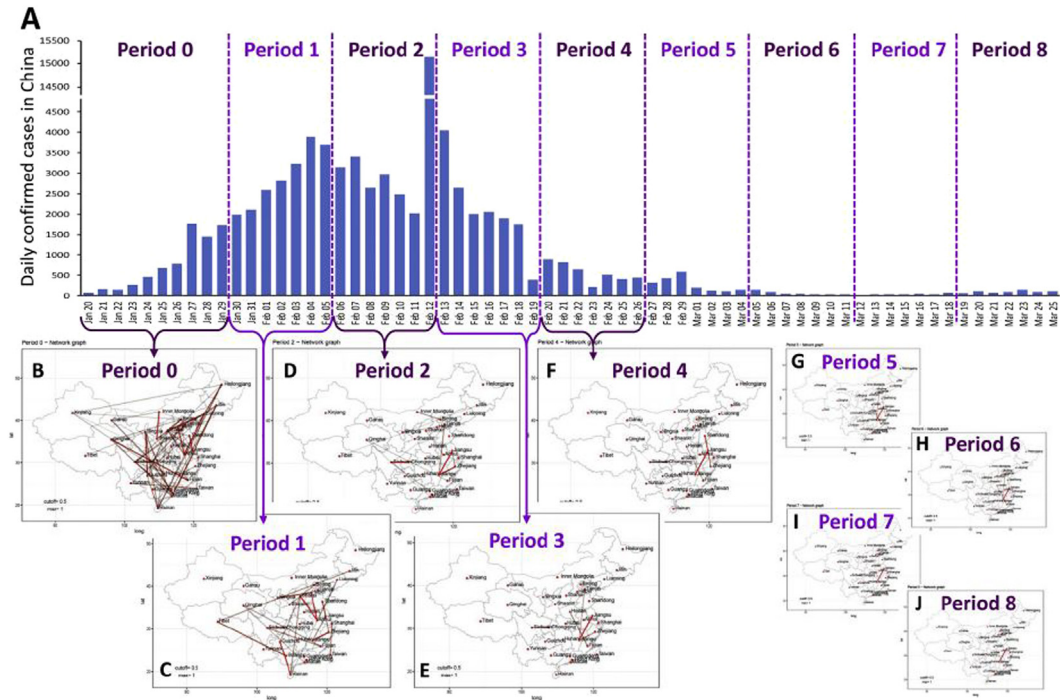


FIGURE 4-1 Data visualization using network graphs representing daily cases in China (So et al., 2020).

using the correlation between different geographic regions for different intervals of time. The graphs are shown in Fig. 4-2.

By analyzing the network graphs, they have also identified that if two different regions are less frequently connected, the density of the lines in the graph decreases, which in turn decreases the tendency of coevaluation of virus spread amid these regions. The researchers have also analyzed the network graphs during the lockdown period and have identified that there was a decrease in the connectedness among the regions, which provides an early sign of an improvement in the epidemic. Similarly, graphs for other countries, namely Europe and United States, were analyzed for different periods. During the research study, the authors demonstrated that network analysis is a powerful approach to provide a clear visualization of epidemic risk through a degree of connectedness in network graphs. Network analysis can enhance traditional modeling techniques for enhancing the approximation of epidemic risk and provide more appropriate indications to take proactive measures to minimize the spread of the virus.

Pradipta et al. have used automatic phase detection techniques to visualize and compare the spread rate of the COVID-19 epidemic across different countries over time (Biswas et al., 2020). The authors have surveyed existing visualization techniques that have been employed on several websites and different media outlets. They have investigated existing data

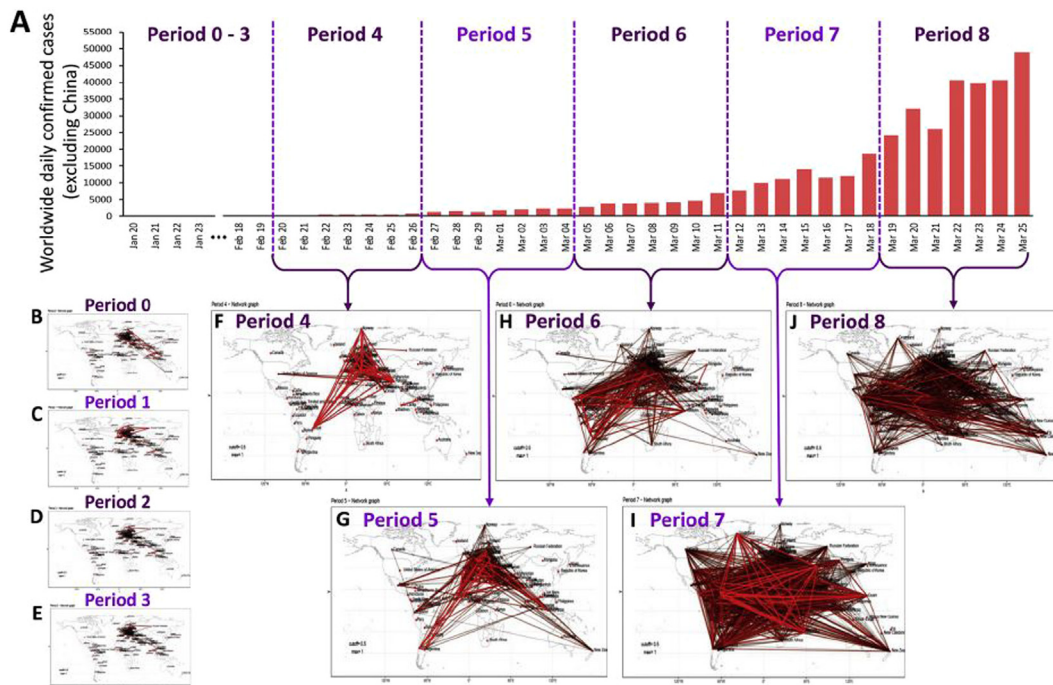


FIGURE 4-2 Data visualization using network graphs for daily confirmed cases worldwide (So et al., 2020).

visualization techniques for representing the enduring spread of the epidemic and have proposed a novel web-based interactive tool. The tool is capable of automatically distributing the period of virus spread based on the increasing rate of COVID-19 cases. The tool uses a knee-detection algorithm that divides exponential spread into several linear components. The algorithm progressively divides the time duration. The authors have constructed different graphs such as linear, exponential, and parabolic graphs and have compared them at different stages for different countries. The novel thing the authors have proposed is that instead of splitting the timeline uniformly for all countries, they have plotted the stage of virus transmission in a data-centric fashion for every country. Their tool has been deployed on a website and can be used for comparative analysis across several countries. The tool is capable of automatically distinguishing and comparing several phases of virus spread across several countries. Their tool can be used to prevent virus spread based on the phases from other countries. In their research, they have used automatic phase detection to divide the duration of virus spread and disease based on the increased rate of new cases using the knee detection algorithm. They have also used a combination of line and bar graphs to indicate the current virus trend. Fig. 4-3 depicts a snapshot for the total number of COVID-19 cases versus deaths and also for the number of new cases versus deaths per day (Biswas et al., 2020). The graph depicts the spread of the virus over different periods. From the line

graph, we can interpret that the total number of deaths and the total number of COVID-19 cases have increased from the period of January to April, resulting in an exponential curve. Thus line graphs help the user in analyzing the spread of the virus over different time zones.

Pradipta et al. have also used a combination of the pie chart and line graph to compare the diverse phases of spread for each country. Fig. 4–4 depicts data visualization using line graphs and pie charts across multiple regions where a plot of the number of new cases versus the number of deaths due to COVID-19 per day has been presented (Biswas et al., 2020).

Additionally, they have compared statistics for individual countries or states based on the count of COVID-19 cases and deaths, the count of new cases and daily deaths, and comparison across different regions based on the count of new cases and the number of deaths per day using three separate graphs. The authors have analyzed the impact of the virus using

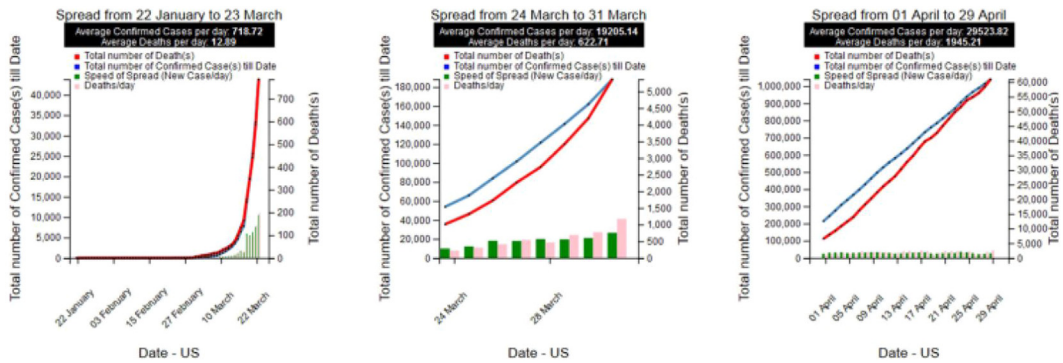


FIGURE 4–3 Data visualization using graphs that combine line and bar charts (Biswas et al., 2020).

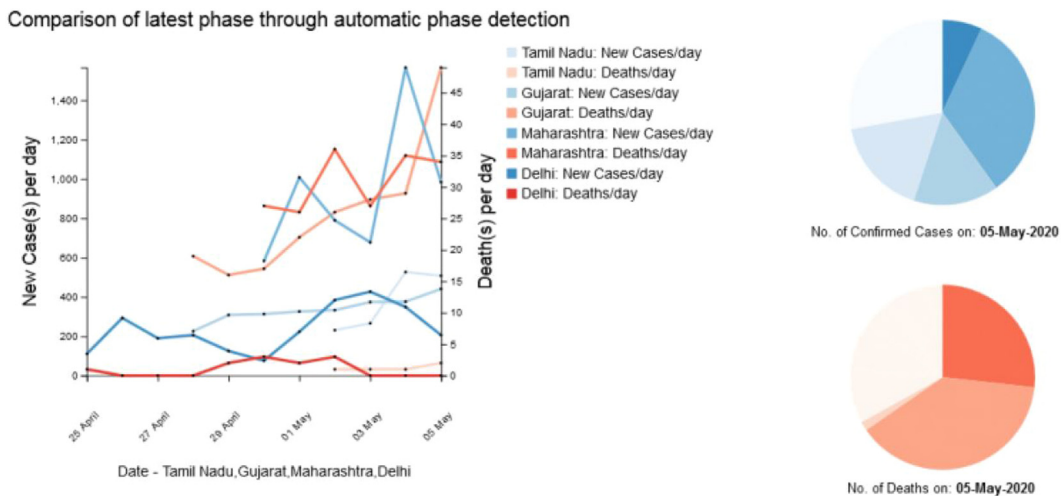


FIGURE 4–4 Data visualization using line graphs and pie charts (So et al., 2020).

radar charts for different countries where the count of new COVID-19 cases per day was greater than 850. The highlight of the proposed work is that the system is capable of indicating the rate as well as the trend of virus spread over time by linking them with past instances. The system is also capable of forecasting the future rate of virus spread.

Weng et al. have analyzed and visualized COVID-19 data using digital earth spatial-temporal framework (Weng et al., 2012). In their study, they have employed the China Star Epidemic tool for data analysis and visualization, which is capable of performing complete and detailed analysis and visualization of COVID-19 data. The tool integrates many attributes of pandemic data into a single system. The highlighting feature of their proposed work is that digital earth can be used for visualizing the varying pattern of pandemic data in both the time and space dimensions. The authors have used spatial-temporal epidemiological analysis that comprises irregular lattice data, case event data, geostatistical data, and regular lattice data. They state that the epidemic data can be analyzed in three dimensions, namely case info, time, and space. They have also used resolution remote sensing data to get visualization hands-on spatial-temporal pandemic data. They employed a Universal Transverse Mercator (UTM) grid for building the index of remote sensing data. They have used three methods for processing epidemic data: (1) spatial-temporal data exploration, (2) spatial-temporal data mining, and (3) spatial-temporal data visualization. They have also implemented 'Thinking Spatially,' which is one method of thinking and is a group of cognitive skills.

Maciejewski et al. created a tool known as PanViz to analyze the outcome of decision measures implemented during the virtual influenza pandemic scenario (Maciejewski et al., 2011). The authors have considered spread vectors that were based on the virus' origin and the distance traveled over time. Their tool helps in analyzing, modeling, and exploring epidemic influenza. The interface employs linked views to display statistical data about the population under analysis. It supports filter controls for demographic data and age. The parameters of the proposed model, namely spread origin and mortality rate, are modifiable via the graphical user interface (GUI). Their work emphasizes the control and containment of pandemic outbreaks. They have displayed georeferenced data using maps with a scrolling facility for the users.

Muthusami et al. performed statistical analysis and visualization for the reported cases of COVID-19 (Muthusami and Saritha, 2020). They also explored time-series data using different visual data analysis tactics with R programming language to provide a clear outcome of the outbreak. The authors have created graphs to give awareness of the virus spread. Additionally, regression and generalized linear models of data constructed from COVID-19 time-series data have been analyzed for infected, recovered, and death cases. Fitted models gave better statistical results during the analysis. Later, based on the obtained results, the authors demonstrated that their results were statistically significant. The authors also applied statistical analysis on the dataset of Spain. They have also analyzed the number of deaths for different countries using linear and exponential curves. Line graphs were employed to display the number of confirmed cases, deaths, and recovered cases across several countries.

Following are the key points based on the data visualization techniques that have been employed by various researchers.

1. Line graphs have been used to show the changes of pandemic spread over time. For instance, line graphs have been used to explain an increase or decrease in the number of COVID-19 cases for different periods. They have been employed for tracking changes in virus spread over short and long periods.
2. Bar graphs have been employed to compare the different sets of data based on different categories. For instance, researchers have used bar graphs to plot the number of deaths for different age groups. The graphs have also been used to show the progress of the pandemic disease based on the daily new cases. These graphs can depict meaningful trends of pandemic diseases at a glance. For instance, one can identify whether there is an increase in new cases or the number of new cases has stabilized, or even might have begun to drift downwards.
3. Pie charts are used to depict the relative sizes of data. They display the data using a circular graph. The complete pie chart represents 100% of a whole and the portions of the pie represent fractions of the whole for every category. Based on the percentage of every pie, researchers have employed them for presenting data related to pandemics. For instance, these have been used to depict deaths by the COVID-19 virus based on age group, ethnicity, etc.

To summarize, this section included various data visualization techniques to analyze the spread of the virus across different countries. The following section describes mathematical models that have been put to use to analyze and predict pandemics.

4.4 Epidemic modeling for analysis and prediction of pandemics

Although vaccines for several epidemic diseases are available, these diseases still cause a lot of distress and mortality across the globe, predominantly in developing nations. In these nations, severe diseases like heart stroke, diabetes, cancer, arthritis, and heart conditions have also received a lot of attention in comparison to epidemic diseases, as they are the primary reasons for death or disability. To study any epidemic, it is imperative to understand its transmission mechanism. For most infectious diseases, the transmission of the pathogen from healthy people to infection is known. People are also familiar with the infection spread mechanism. However, it is very challenging to understand the transmission interactions in a large community; it is very hard to find the large-scale dynamics of disease spread. In such a scenario, the structure of mathematical models can help in analysis.

Although, real-time modeling of the infectious diseases that have spread to the human population is always impossible or expensive. An epidemiological model uses a microscopic description of an infectious disease. In many disciplines, it is possible to conduct experiments to obtain information and its related test hypotheses. Datasets can be taken from the natural occurrence of epidemics. However, data are often incomplete due to underreporting. Due to the lack of reliable data, it becomes difficult to make accurate parameter estimations. However, mathematical modeling and computer simulations can be used to compute approximate parameter value ranges for the available data sets. It is important to examine

the epidemic models corresponding to the infectious diseases we are trying to simulate, as each of them have both limitations and capabilities that must be validated. Sometimes, queries cannot be addressed using epidemiological models, but the investigator has to look for the right combination of available data, a curious problem, and an appropriate model for a correct solution to a problem. The epidemiological model parameters should be clearly defined like whether a parameter depicts the transmission rate, contact rate, or infection duration. The model can be used to find many quantitative estimates and predict the spread of disease. The model can also be used to determine the impact of the parameter change on predictions, that is, how sensitive predictions are related to a change in parameter values. During the investigation, the best estimation of parameters can be done. The best estimates mostly impact the predictions. It is possible to find the best estimates of the parameters of an epidemic model using an index known as the basic reproductive number (R_0). R_0 is a very useful metric for epidemic analysis. The reproductive number helps in the quantification of virus transmission. The reproductive number is defined as the ratio of the transmission rate (β) and recovery rate (γ). It signifies how many individuals get infected by an infected person during the infectivity period. [Table 4–2](#) shows the R_0 of various epidemics that have occurred in past ([LePan, 2021](#)). It can be observed from [Table 4–2](#) that measles has the highest value of R_0 among all, as it is the most contagious, with R_0 ranging from 12 to 18. This means that, on average, 12–18 individuals are infected by an infected person. The 12–18 individuals that get infected are susceptible populations and believed to be unvaccinated.

It is possible to generate an epidemic curve using a simple epidemic model. It is also possible to forecast the fraction of the total population that will get infected during each day of an epidemic. Besides, analytical tools help in drawing inferences from the model. An important conclusion about the reproductive index, which helps in analyzing the epidemic growth, is as follows: if R_0 is less than 1, it means the disease will not spread quickly. It results in a decay in the number of infected individuals. On the other hand, if R_0 is greater than 1, it means that infection will spread and the infected population will grow with time. Whenever an epidemic occurs, its size does not depend upon the number of infected people; rather, it depends upon the initial proportion of susceptibility, $S(0)$, and R_0 . The final epidemic size,

Table 4–2 Basic Reproductive Number (R_0) of diseases.

Disease	R_0
MERS	0.8
Influenza	1.5
Ebola	2
Covid-19	2.5
SARS	3.5
Mumps	4.5
Rubella	6
Smallpox	6
Measles	16

that is, the proportion of the infected population, will always be sensitive to the initial proportion of the population and will always be smaller than $S(0)$.

These inferences, insofar as they apply in reality, have some critical results. If we vaccinate a certain proportion of the population before the onset of infection so that the initial population of susceptible people can be reduced to $S(0) < \gamma / \tau$, we can prevent an epidemic. This result is based on the notion of herd immunity, which can be attained if a large portion of the population is vaccinated to prevent an epidemic. If we do not get high enough vaccination coverage, we would have only reduced the size of the epidemic and not prevented it. $S(0) < \gamma / \tau$ condition. Other ways to achieve this condition and eliminate an epidemic are as follows: (1) reducing the transmission parameter by isolating measures of infection or social distancing or imposing lockdown; and (2) the treatment of infectious diseases, which thus increases the recovery rate.

The novel coronavirus (COVID-19) pandemic is an emerging disease that has grabbed worldwide attention. It has affected almost all the major countries around the globe. It is important to model such a disease correctly to identify its spread. In this regard, predictive mathematical models play an important role in investigating the epidemic spread due to the lack of ineffective vaccines. The spread of any virus is influenced by various factors such as the size of the population, the number of infectious days, the recovery period, the number of people with whom an infected person comes in contact, and the probability of contracting the virus from an infected person. Mathematical models can aid in understanding the risks associated with the spread of pandemics of potential infectious diseases. For example, the models can predict the chance that an infection will invade a particular nation, the expected number of cases within a specific timeframe, or the social norms' effectiveness. Epidemic models are deterministic models for such epidemics. These simple models are formulated as initial value problems for ordinary differential equation systems and are analyzed mathematically. The models mimic reality to provide useful outputs. Models such as SIR, SIS, SIER, and SIR-L have been used traditionally to study and analyze the spread of an epidemic. Many variants of the SIR model are also used by researchers to study epidemics. In the following section the basic epidemic models adopted widely are described.

4.4.1 The susceptible-infected-recovered model

The SIR model is a simple epidemic model. It can be used to study the spread of an infectious disease in a large population. It was developed by Kermack and McKendrick (Kermack and McKendrick, 1927). This model assumes that the population is divided into three compartments: S (susceptible), I (infected), and R (recovered). It is assumed that the latency period for the infection is zero and no birth and death occur during the modeling period. S, I, and R are the functions of time t . They change according to a system of differential equations. The susceptible compartment has a population that is healthy and has a high probability of getting infected. The infected compartment represents the number of infected people. The infected people have the disease and can transmit it to healthy people (susceptible). The recovered compartment represents the number of people who had the

disease in the past and they have recovered. The R individual cannot become infected in the future and they cannot transmit the disease to others. They will have natural immunity to the disease.

New infections occur due to contact between infectious and susceptible people. In the SIR model, the rate at which new infections occur is βSI , where β represents the transmission rate, which represents the fraction of susceptible that can be infected postcontact with the infected person. Whenever a new infection occurs, the infected person moves from the susceptible compartment to the infected compartment. Besides, no person will enter or leave the susceptible compartment. Here is the first differential equation that shows the rate of change in the number of susceptible people:

$$dS/dt = -\beta SI \quad (4.1)$$

The infected person, whenever they recover from the disease, enters the recovered compartment. The rate of change of recovery is given in Eq. (4.3), where γ is the recovery coefficient that depends upon the duration of infection and the total number of infected people. The rate of change of infection rate depends upon the change in susceptible and recovered people and that has been depicted in Eq. (4.2). Thus we have our other two differential equations:

$$dI/dt = \beta SI - \gamma I \quad (4.2)$$

$$dR/dt = \gamma I \quad (4.3)$$

Since we assume the population in a closed system, the sum of $S + I + R$ remains constant.

4.4.2 The susceptible-infectious-susceptible model

This is a variant of the SIR model where an infected person who recovers does not confer immunity from the infection. Thus postrecovery, the person becomes susceptible. Infections like influenza, cold, and flu are studied using the SIS model. SIS models the interaction among people in two compartments: susceptible (S) and infective (I). The susceptible (S) compartment is the proportion of the population who are prone to get infected or those who have recovered from infection in the past. The infective (I) compartment represents the people who are infected with the infectious disease and can spread the infection. The SIS model can be described using the differential equation as

$$dS/dt = -\beta SI + \gamma I \quad (4.4)$$

$$dI/dt = \beta SI - \gamma I \quad (4.5)$$

In the past, epidemic diseases such as meningitis, influenza, gonorrhoea, and childhood diseases, whose spreading processes show heterogeneity of infectivity in infectious

individuals, had been studied using the SIS model. It is also beneficial in analyzing waterborne diseases like cholera (Chen et al., 2018).

4.4.3 The susceptible exposed infective recovered model

The SEIR model simulates the time-histories of an epidemic. SEIR models the dynamic interaction of people using four compartments: susceptible (S), exposed (E), infective (I), and recovered (R). The classical SEIR model can be described by a sequence of differential equations as follows:

$$dS/dt = -\beta SI \quad (4.6)$$

$$dE/dt = \beta SI - \sigma E \quad (4.7)$$

$$dI/dt = \sigma E - \gamma I \quad (4.8)$$

$$dR/dt = \gamma I \quad (4.9)$$

Similar to the SIR model, in this model the susceptible (S) compartment represents the fraction of the population that comprises healthy people who are potentially subjected to get an infection. Initially, the whole population is considered to be susceptible. The part of the population that has been infected but does not show symptoms is considered to be exposed (E). At this stage, an infected person may or may not cause infection to others. This can be considered a latent phase. The infective (I) compartment represents the infected population postlatent period. The recovered (R) compartment denotes the fraction of the population that has been healed from the infectious disease. They are generally not reintroduced to the susceptible compartment. It has been assumed that they have become immune to the disease. This aspect is strongly debated, as in the case of COVID-19, a few countries have recorded the second infection of recovered people. In the classical SEIR model, the R category also comprehends the individuals who died of the disease.

The SIR model has been popularly used among researchers to study epidemics such as measles, polio, and flu. Toda et al. used the SIR model to estimate the situation for COVID-19 (Toda, 2020). They observed that the estimated transmission rates, β , are heterogeneous across countries. The transmission rate β was found to be low in China. While in Japan and Korea, the neighboring countries of China, the initial transmission rate β was very high at around 0.2–0.4. It has been found that the heterogeneity in transmission rate β is maybe because early epicenters have already taken mitigation measures. They stated that since the situations and policies were continuously changing, the whole data cannot be used for the estimation. Hence, they took the data for a 14-day window. They also stated that experience or data are not so relevant in the new situations and a mathematical model could help in predicting the scenarios shortly. Wenzhi Chen has established a mathematical model for the Ebolavirus and analyzed data to predict its spread (Chen, 2016). He had proposed a modified SIR model with additional parameters to project the outbreak of Ebola more

efficiently. The paper did not consider the natural birth rate and mortality caused by the number of changes. In the incubation period, it is assumed to be not infectious. He projected the relationship between infection cases and the weeks using the SIR modeling, and the predictions came out to be very accurate, which verifies that SIR modeling is highly efficient in predicting the spread of an epidemic. Along with stating mathematical equations like no-epidemic situations, he has also given a qualitative analysis of the situation and has suggested mitigation policies to control the spread. He has also shown the impact of the virus on the economy and has suggested policies for the revival of the economy of the region. António M. Correia applied the SIR model to forecast the evolution of infectious diseases, namely measles and hepatitis C, in Portugal. They used the Portuguese health system data, collected over the course of 11 years, from 1996 to 2007 (Peixoto et al., 2011). They investigated their results to forecast the progression of those viruses in consecutive years. They obtained the optimal values for the parameters: minimum average error and maximum of the correlation coefficient r for best fit of their model. They inferred that although 11 years' worth of data is scarce, it can still provide us some useful and interesting insights into the growth of measles and hepatitis C virus.

Elena Loli Piccolomini proposed a Susceptible-Infected-Exposed-Recovered-Dead (SEIRD) model to analyze and predict COVID-19 spread in Italy (Peixoto et al., 2011). They realized the importance of the exposed parameter in their model, and hence shifted to SEIR. To better analyze the spread of disease and the mitigation measures, they moved to a modified SEIR-D model, including the death parameter. To study the restrictions obtruded by the Italian government, they divided the whole period $(0, T)$ into two subintervals, namely $(0, t_0)$ and (t_0, T) , where t_0 corresponds to the time when the restrictions had started to produce a valuable change in the trend of the data. After applying the model, they calibrated the t of SEIR parameters by solving nonlinear least squares problems with positive constraints to optimize the parameters. In their model, the infection rate coefficient was adaptively modeled as an inverse function of time, which took into account the Italian government's restrictions. The results obtained by fitting the models to the different regions showed a very good fit to the data.

4.5 Machine learning approaches for analysis and prediction of pandemics

ML has been immensely used in identifying patterns in many application areas including education, finance, healthcare, security, and surveillance. Its applications can also facilitate the management of epidemics.

Due to the temporal nature of an epidemic, the problem generally fits in the category of time-series prediction. The traditional autoregressive (AR) methods have been extensively used to study the spatial-temporal epidemic data (Ibrahim et al., 2017). The basic principle is to combine the past data points linearly to model the future state. Another method is Gaussian Process Regression (GPR), which utilizes a nonlinear kernel for modeling complex

temporal patterns (Perrotta et al., 2017). These methods extend the prediction power. These methods are simple as they require a smaller number of parameters for epidemic prediction, but they are also limited in expressiveness. Statistical methodologies are also commonly used approaches to study epidemiology (Greenland and Poole, 2013). The statistical methods and multivariable regression techniques are based on testing the hypothesis. These techniques are used to define the association between independent and dependent variables. The independent variables represent the predictor, whereas the dependent variables are the outcome variables. Using these methods, a dependent variable is predicted (outcome) based on multiple independent variables (predictor). Using these methods an epidemic's treatment effects can be studied effectively. Although these modeling techniques provide useful insights for investigators and practitioners, these methods are based on varied assumptions that may be unrealistic (Breiman, 2001).

ML is a data-driven technique based on mining and learning from the data to derive useful insights. This has been a favorite choice among researchers to study epidemiology. ML can be used for health predictive analysis to analyze the spread of infectious diseases so that preventive measures can be taken (Ravi et al., 2017). Fig. 4–5 shows the steps involved in the ML technique.

Data acquisition is the initial step in ML as every epidemic has its unique features like its epidemic contagiousness intensity and method of transmission. Fig. 4–6 depicts epidemics' dissemination factors in health predictive analysis (Ibrahim et al., 2017). The epidemic disease dissemination occurs due to varied means, namely physical network, geographical location, clinical studies, and social media. The health analysis uses ML methods for making a prediction. Harmful viruses cause the spread of infectious diseases. A physical network is a medium through which the virus propagates among hosts. Population density and hotspots are the two factors that form the physical network. The population density is the fraction of the population in a region. A higher population density results in an increase in the virus spread, which eventually can cause an epidemic outbreak. The hotspot is an area where a lot of physical contact is present and high ingress and egress exist among people. Places such as amusement parks, shopping malls, tourist spots, and flight connecting points are considered to be hotspots.

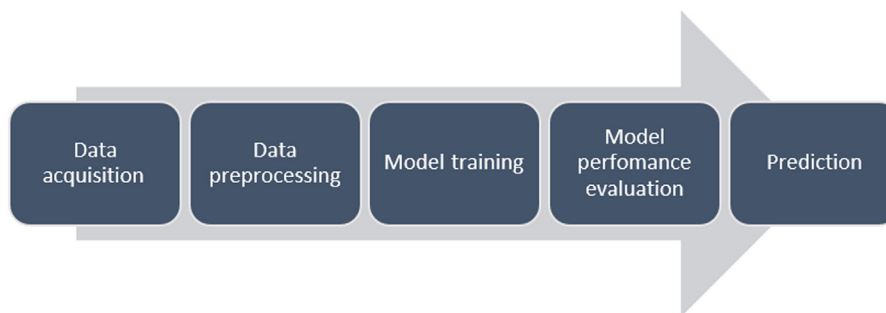


FIGURE 4–5 Machine learning process.

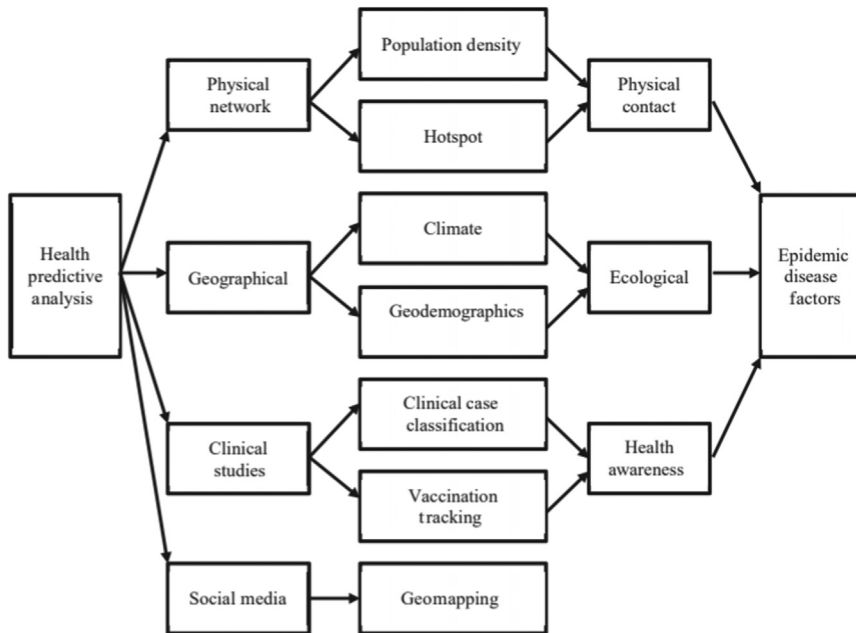


FIGURE 4-6 Health predictive analysis and correlation of epidemic disease dissemination factors.

The climate and geodemographics are considered to determine a geographical location. Due to deforestation and desertification, the climate has been largely affected. The variation in the climate has promoted the productivity of air- or water-borne diseases spread through insects, such as malaria and dengue. These are believed to be created due to man-made disasters. Geodemographics is also one of the contributing factors to epidemics. In geodemographics the population within a community is classified in terms of the number of families, their sanitation level, the neighboring economic activities and literacy level, etc. (Ravi et al., 2017). The poor management of waste disposal or deforestation in a geographic location can nurture an epidemic disease outbreak. Epidemic disease tracking using clinical studies can be done based on clinical case analysis and tracking the vaccination. Clinical studies can classify or cluster the patients by correlating their electronic health records, which consist of implemented and accepted procedures for treatment defined by developed nations. This clinical study can also help in vaccination tracking (VT). VT is used to find the vaccine type utilized and the maximum consumed vaccine in an infected area. Social media can cater to enriching health predictive analysis of epidemic diseases (Ibrahim et al., 2017). Social media is a popularly used medium for information exchange. The social media data in the form of tweets and posts can help in geomapping the disease spread (Ibrahim et al., 2017; Perrotta et al., 2017). The words' classification within the posts and tweets can be grouped with the geotagging, which can help in personalization. Besides, it can be used to

flag an increase in the number of symptoms in a particular area and to determine the infected area. BlueDot, a Canadian startup, was able to first spot symptoms of the coronavirus by processing the data residing on digital media and official public health organizations and global airline ticketing data (Bogoch et al., 2020). All these mentioned factors can help in disease identification, detection, prediction, and control via feature classification.

Postdata acquisition, the preprocessing steps of the epidemic disease data are applied to prepare the data for the ML model. Data preprocessing steps include data cleaning, data integration, data transformation, and feature selection. Once the right kind of data is selected, then an appropriate ML model is chosen for model building. Models are built using ML techniques. ML techniques can be used for data classification and regression problems. Classification models are generally considered for categorical labels. Regression models are considered for continuous labels. It is imperative to normalize the continuous data before training a model to standardize the scale of multiple continuous features and enhance the computational performance. The selected ML model is then trained using the training set of the data. This involves the model making iterative runs through the data while computing a function to classify the data and hence recognize an underlying pattern. For epidemic outbreak prediction, the disease-related features are extracted, analyzed, and correlated for classification. After the model is built using training data, its performance is evaluated on testing data. There are various metrics available to evaluate the performance of an ML model: recall, precision, and F1 score. This step also ensures that the model is not an “over-fit” or an “under-fit” model. In the end the prediction is done to classify the outbreak factors or to classify whether a patient is likely to contract a particular virus. Predicting new trends in epidemic data is an important and interesting research problem that has attracted the attention of many medical practitioners and researchers in the ML community.

Machine learning has been applied to analyze the web search and geolocation data related to restaurants to identify possibly unsafe restaurants (Sadilek et al., 2018). ML approaches have been largely employed to fight the COVID-19 pandemic (Kushwaha et al., 2020). For classification problems in a health-related domain, generally, data is skewed or imbalanced. Imbalance occurs when one class label is more prevalent than another. In such cases the prediction accuracy might be high, but that too happens due to majority class labels. Techniques like undersampling, SMOTE (synthetic minority oversampling technique), and ensemble-based techniques are popularly used in many application areas of epidemics such as the cancer survivorship prediction of mesothelioma patients (Fotouhi et al., 2019; Chicco and Rovelli, 2019).

Furthermore, other categories of ML algorithms such as deep learning, which is used to describe artificial neural networks (ANNs) with many hidden layers, have been used extensively to detect viruses from medical imaging data such as coronavirus detection from a high-resolution CT scan (Kaggle, 2021a) and screening of the coronavirus disease in pulmonary CT scan images (Chen et al., 2020; Xu et al., 2020). These models are very complex and provide better accuracy. The problem in these approaches is in hyperparameter selection and training. They are very computationally intensive.

4.6 Semantic technologies

A large number of scientific journals and patents have been published in the past decades worldwide, constituting essential resources for supporting drug research as well as helping in the expansion of pioneering methods to fight against novel ailments and viruses (Aras and Ph, 2021). Though patents related to a new pandemic like COVID-19 will take some time to be submitted and approved; previously published research articles and patents might already contain essential information related to antiviral drugs, vaccines, medical treatment, etc. For most of the cases, relevant data is disclosed in patent documents, which are published approximately 18 months after they are filled. This fact signifies that the patents related to COVID-19 may appear from March 2021. Thus finding and exploring knowledge in patents is a cumbersome task as it requires a lot of manual effort to sift through several documents and technical descriptions to extract relevant information (Aras and Ph, 2021).

Semantic techniques help in validating data as well as in governing rules for storing as well as manipulating the data when required. The data is conveyed with a suitable meaning such that it is well understood and deprived of any difficulty (Rawal et al., 2021). Rule-based analysis techniques have been improved in recent years because of the increased use of text mining, natural language processing, and ML approaches. It has been identified that the inclusion of semantic information using implicit semantics, that is, through word embedding or using explicit domain information such as expertly created ontologies, can effectively help in exploring data collections. Semantic techniques can also be employed to inquire and answer complex normal language questions in a very interactive manner. This makes it feasible for both collecting factual information and exploring, justifying, and identifying further derived substantial information within the text by writing complex semantic queries. These related and subsequent queries can be easily linked with the user's natural language questions which are associated with the important aspects of research questions. During the current COVID-19 pandemic, Kaggle posted a call for data scientists to develop innovative methods that help in answering a list of research questions that are related to COVID-19. A research dataset called CORD-19 (COVID-19 Open Research Dataset) has been made publicly available and is updated periodically by the White House and an alliance of leading research organizations (Aras and Ph, 2021).

Scientists apply existing knowledge to gain new insights and find answers for complex research queries. Their search is based on finding out answers for interdisciplinary. Thus the informational requirements for them can be quite complex, as their questions, as well as search queries, must be articulated in a meaningful way to be coherent to other researchers working in similar fields. Therefore keyword-based statistical techniques, as well as semantic informatic, must be considered for representing and seeking valuable information to satisfy particular information needs.

Aras developed a data processing and analysis pipeline that applies natural learning processing as well ML techniques for text analysis of scientific articles and patents. A semantic search engine was integrated into the data analysis pipeline. The engine is constructed on a patent corpus that has been created after extracting relevant documents from full-text

databases of STN[®]. Their semantic engine is capable of identifying and extracting the most important sentences as well as passages from patent texts to answer specific questions for users. They have also performed an embedding model for capturing the semantics of words or phrases using deep learning approaches. They have also processed data for similarity analysis by generating a matching similarity index between a user query and sentences in the full text from scientific articles or patents. The researchers have used an approximate nearest neighbor algorithm to determine the most pertinent as well as semantically similar paragraphs and sentences for a given query vector.

Rawal et al. also analyzed the COVID-19 virus pattern that was based on the semantic-web approach with the incorporation of description logic, which elucidates statements based on reasoning (Rawal et al., 2021). They combined description logic with the semantics to give semantically valuable meaning to the disease pattern of the novel coronavirus. They have used description logic ontologies along with web-based semantic techniques to remove ambiguity that was developed due to the spread of coronavirus. They have demonstrated that semantics, when combined with description logic, imparts meaning to words such that their clarification is done correctly.

Thus it can be observed that the semantic-driven searching techniques can be employed for users to explore information from patents and research articles and gain insight into selected topics. Semantic techniques can be used to extract biomedical entities which can be linked to knowledge graphs for analyzing viruses and infections. For instance, bio-medical entities from scientific journals and patents can be linked to CovidGraph (Aras and Ph, 2021). This allows real-time investigation of various sources such as drugs and their chemical information, research articles, and scientific journals. Additionally, scientists can also obtain information for developing new antiviruses or new drugs that are based on active ingredients that are effective against SARS-CoV, RNA virus. They can also extract the chemical structures of drugs that are mentioned in published articles to develop new antiviruses using semantic techniques.

4.7 Summary

An epidemic is not a favorable situation for anybody, as society confronts the challenges of public health and economics in the wake of any pandemic. The pandemic, despite all the harm it does to society, also raises many relevant questions on the approach to future epidemics and how the current state of relative helplessness can be overcome. The spread of a pandemic and the public's desire for information has led to the creation of many open-source datasets and visualization mechanisms that can be used to perform pandemic disaster analytics. The analysis and study of data from several sources help in interpretation and pandemic-related research. The geospatial data of a region help in analyzing the epidemic outbreak. This chapter has comprehensively reviewed the available datasets, visualization techniques, and different available approaches to conduct a study on any pandemic or epidemic. The approaches that are gaining popularity are epidemic models, ML models, and semantic technologies.

Epidemic models such as SIR, SIER, and SIR-L have been extensively employed to forecast the spread or occurrence of a disease. Many variants of the SIR model are also used by researchers to study epidemics. Using the epidemic modeling concepts, researchers can predict the pandemic's curve and how severe it could turn in the future. These models can also be employed to discover the sensitivity of predictions of changes in the parameter values.

ML approaches have been popular among medical researchers for the past two decades. These methods have contributed toward epidemic handling in many dimensions. Studies performed using ML on past epidemics have encouraged researchers by giving a new angle to analyze the nature of the virus. In addition, it had helped in the identification of appropriate drugs, identification of patients at high risk, identification of severe infection symptoms, or selecting useful features for the prediction of confirmed cases. Although classic, statistical techniques, modeling can provide satisfactory models but it can likewise fail to comprehend the details contained within the data.

Studies have revealed that semantic techniques make it feasible for users to explore relevant information from patents and scientific journals in order to gain insights into selected topics and techniques, thereby allowing the users to articulate their information needs using a few words.

Review questions

1. How are epidemics affecting lives?
2. Why is the predictive problem most crucial?
3. What are the techniques to model the epidemic behavior?
4. How do we get the best estimation of parameters that impact the epidemic model?
5. What are the commonly used visualization graphs for studying the behavior of any pathogen?
6. What are the importance of epidemic models, ML models, and semantic techniques in studying any infectious disease?
7. What are the openly available data repositories for studying epidemics?
8. How can mathematical models help in modeling and analyzing pandemics?

Discussion questions

1. How does a pandemic outbreak impact the social, well-being, and economy of a country?
2. What are the operational concerns for different sectors such as healthcare, information technology, finance, energy, agriculture, mining, manufacturing, real estate, and transportation in case of a pandemic?
3. How can the crisis management team operate effectively to minimize the spread of deadly viruses or diseases?

4. How can traveling employees be brought back to their homes if they get infected by the virus?
5. Describe the assurances that need to be provided to the facilitators working in different sectors so that they are safe at their workplace.

Problem statement

How do we educate people, employees, and facilitators in a pandemic situation so as to minimize its impact on the economy of a country?

References

- Aras H. and D. Ph, 2021. Applying Semantic Technologies For Knowledge Discovery From Patents – A Case Study For The COVID-19 PANDEMIC, pp. 1–6.
- Biswas, P., et al., 2020. COVID-19 data visualization through automatic phase detection. *Digit. Gov. Res. Pract.* 1 (4), 1–8. Available from: <https://doi.org/10.1145/3411756>.
- Bogoch, I.I., Watts, A., Thomas-Bachli, A., Huber, C., Kraemer, M.U.G., Khan, K., 2020. Potential for global spread of a novel coronavirus from China. *J. Travel. Med.* 27 (2), 1–3. Available from: <https://doi.org/10.1093/jtm/taaa011>.
- Breiman, L., 2001. Statistical modeling: the two cultures. *Stat. Sci.* 16 (3), 199–215. Available from: <https://doi.org/10.1214/ss/1009213726>.
- Car, Z., Baressi Šegota, S., Anđelić, N., Lorencin, I., Mrzljak, V., 2020. Modeling the spread of COVID-19 infection using a multilayer perceptron. *Comput. Math. Methods Med.* 2020. Available from: <https://doi.org/10.1155/2020/5714714>.
- Chan, M., 2010. *World Now at the Start of 2009 Influenza Pandemic*. WHO.
- Chanprasopchai, P., Tang, I.M., Pongsumpun, P., 2018. SIR Model for dengue disease with effect of dengue vaccination. *Comput. Math. Methods Med.* 2018. Available from: <https://doi.org/10.1155/2018/9861572>.
- Chen W., 2016. A mathematical model of Ebola virus based on SIR model. In: *Proceedings of the International Conference on Industrial Informatics—Computing Technology, Intelligent Technology, Industrial Information Integration*, vol. 128, pp. 213–216. <https://doi.org/10.1109/ICII.2015.135>.
- Chen, S., Small, M., Tao, Y., Fu, X., 2018. Transmission dynamics of an SIS model with age structure on heterogeneous networks. *Bull. Math. Biol.* 80 (8), 2049–2087. Available from: <https://doi.org/10.1007/s11538-018-0445-z>.
- Chen, J., et al., 2020. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. *medRxiv*. Available from: <https://doi.org/10.1101/2020.02.25.20021568>.
- Chicco, D., Rovelli, C., 2019. Computational prediction of diagnosis and feature selection on mesothelioma patient health records. *PLoS One* 14 (1), 1–28. Available from: <https://doi.org/10.1371/journal.pone.0208737>.
- Data World (2020). Business data on data.world. data.world. <<https://data.world/datasets/influenza>> (accessed 25.01.21).
- Fotouhi, S., Asadi, S., Kattan, M.W., 2019. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *J. Biomed. Inform.* 90, 103089. Available from: <https://doi.org/10.1016/j.jbi.2018.12.003>.

- Greenland, S., Poole, C., 2013. Living with P values: resurrecting a bayesian perspective on frequentist statistics. *Epidemiology* 24 (1), 62–68. Available from: <https://doi.org/10.1097/EDE.0b013e3182785741>.
- Ibrahim N., N.S.M. Akhir, and F.H. Hassan, 2017. Predictive analysis effectiveness in determining the epidemic disease infected area. In: *Proceedings of the AIP Conference*, vol. 1891, no. October 2017 <https://doi.org/10.1063/1.5005397>.
- Johns Hopkins University, 2020. COVID-19 Map. Johns Hopkins Coronavirus Resource Center. <<https://coronavirus.jhu.edu/map.html>> (accessed 25.01.21).
- Johns Hopkins University, 2021. GitHub - CSSEGISandData/COVID-19: novel coronavirus (COVID-19) cases. <https://github.com/CSSEGISandData/COVID-19> (accessed 25.01.21).
- Kaggle, 2021a. Ebola | 2014-2016 | Western Africa Ebola outbreak. <<https://www.kaggle.com/imdevskp/ebola-outbreak-20142016-complete-dataset>> (accessed 30.01.21).
- Kaggle, 2021b. HIV AIDS dataset. <<https://www.kaggle.com/imdevskp/hiv-aids-dataset>> (accessed 30.01.21).
- Kaggle, 2021c. Pandemic 2009 H1N1 swine flu influenza A dataset. <<https://www.kaggle.com/de5d5fe61f-caa6ad7a66/pandemic-2009-h1n1-swine-flu-influenza-a-dataset>> (accessed 25.01. 21).
- Kaggle, 2021d. Zika virus epidemic. <<https://www.kaggle.com/cdc/zika-virus-epidemic/>> (accessed 25.01.21).
- Kermack A., W.O., McKendrick, “A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and physical Character*, Vol. 115, No. 772,,” *Society*, vol. 115, no. 772, pp. 700–721, 1927.
- Kushwaha, S., et al., 2020. Significant applications of machine learning for COVID-19 pandemic. *J. Ind. Integr. Manage.* 05 (04), 453–479. Available from: <https://doi.org/10.1142/s2424862220500268>.
- LePan N., 2021. A visual history of pandemics. <<https://www.weforum.org/agenda/2020/03/a-visual-history-of-pandemics>> (accessed 25.01.21).
- Maciejewski, R., et al., 2011. A pandemic influenza modeling and visualization tool. *J. Vis. Lang. Comput.* 22 (4), 268–278. Available from: <https://doi.org/10.1016/j.jvlc.2011.04.002>.
- Muthusami, R., Saritha, K., 2020. Statistical analysis and visualization of the potential cases of pandemic coronavirus. *VirusDisease* 31 (2), 204–208. Available from: <https://doi.org/10.1007/s13337-020-00610-1>.
- Peixoto M.M., A.A. Pinto, D.A. Rand, 2011. Dynamics, games and science II: DYNA 2008, in Honor of Mauricio Peixoto and David Rand, University of Minho, Braga, Portugal, September 8–12, 2008, Springer *Proceedings in Mathematics*, vol. 2, no. doi: 10.1007/978-3-642–14788-3.
- Perrotta D., M. Tizzoni, and D. Paolotti, 2017. Using participatory web-based surveillance data to improve seasonal influenza forecasting in Italy. In: *Proceedings of the Twenty-Sixth International World Wide Web Conference*, pp. 303–310, doi: 10.1145/3038912.3052670.
- Ravi, D., et al., 2017. Deep Learning for Health Informatics. *IEEE J. Biomed. Heal. Inform.* 21 (1), 4–21. Available from: <https://doi.org/10.1109/JBHI.2016.2636665>.
- Rawal R., K. Goel, C. Gupta, 2021. COVID-19: disease pattern study based on semantic-web approach using description logic, pp. 1–5, doi: 10.1109/inocon50539.2020.9298278.
- Sadilek, A., et al., 2018. Machine-learned epidemiology: real-time detection of foodborne illness at scale. arXiv . Available from: <https://doi.org/10.1038/s41746-018-0045-1>no. July.
- So, M.K.P., Tiwari, A., Chu, A.M.Y., Tsang, J.T.Y., Chan, J.N.L., 2020. Visualizing COVID-19 pandemic risk through network connectedness. *Int. J. Infect. Dis.* 96, 558–561. Available from: <https://doi.org/10.1016/j.ijid.2020.05.011>.
- Toda, A.A., 2020. Susceptible-infected-recovered (SIR) dynamics of COVID-19 and economic impact. arXiv 1–15.
- Union E., 2021. HIV infection and AIDS data - datasets. <<https://data.europa.eu/euodp/en/data/dataset/hiv-infection-and-aids-data>> (accessed 30.01.21).

Weng J., Y. Xu, A.R. Sharma, 2012. Epidemic analysis and visualization based on Digital Earth spatio-temporal framework. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, pp. 7220–7223.

World Data Atlas, Zika Dataset-Knoema.

Xu X. *et al.*, 2020. Deep learning system to screen coronavirus disease 2019 pneumonia. *Med. Phys.*, 6, pp. 1122–1129.