Chapter 3
AI-Empowered Data Analytics for Coronavirus Epidemic Monitoring and Control

Governments and authorities knew little about the virus since the emergency of COVID-19 outbreak. The Chinese government upon the discovery of the early patients in Wuhan, informed WHO on 31 December 2019, as pneumonia of unknown causes. Epidemiologists, data scientists and biostatisticians have been working hand-in-hand for a common mission of trying to characterize and understand the characteristics of the infection as well as the virus itself, which is SARS alike. In front of an unknown disease which is so contagious and dangerous, governments and organization of both private and public started an all out approach using latest data analytics and AI algorithms in the hope of knowing more about the virus, so that the disease spread and progression could possibly be predicted. Data analytics and prediction are equally important if not more than the deployment of AI techniques in confronting the virus and assisting the treatment (as those techniques discussed in previous chapter). COVID-19 is an invisible enemy of mankind in microscopic scale. We can only know about its traits and behaviours as a posteriori knowledge from the collected data and statistics like chasing shadow. This chapter introduces and discusses how some of the prominent AI and data analytics examples that crunch over the data during COVID-19, for forecast and insights.

3.1 AI Predicted COVID-19 Outbreak Before It Happened

Was COVID-19 predicted accurately or by chance by AI prior to its arrival? A Canadian company called BlueDot [1] might have an answer, claiming that they had picked up signals and warned the official about the new disease ahead of WHO and CDC. The prediction was said to have come from scientific inference from an AI model by constantly monitoring the changes from various information sources in lieu of watching over the official statistics of daily reported cases.

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. J. Fong et al., *Artificial Intelligence for Coronavirus Outbreak*,
SpringerBriefs in Computational Intelligence,
https://doi.org/10.1007/978-981-15-5936-5_3

The AI algorithm which of course was kept commercially confidential builds an adaptive model which represents the potential of disease outbreak considering over multiple sources of salient information that are relevant to the disease development. These information sources contain tell-tale-signs hidden in massive amount of data, which could be filtered and discovered by data mining techniques. The tell-tale-signs could be but not limited to the following activities, which can be technically obtained from different means from publicly available data—the so-called big data or otherwise:

Sudden change or deviation from the usual flows of data in:

- International/domestic flights to-and-from certain cities
- Direction and intensity of traffic flow
- Increase of procurement of medical supplies
- Consumers' retail purchase patterns
- Movements or relocations of medical and emergency personnel
- Sharp growth of sentiments or certain topics from social media

Just like any disease, a macro-view of symptoms can be observed from people and their behaviours before a confirmed case emerged and officially reported to authority. There is always some time overhead or latency between the start of the first infection and mass infection. This observation period ensures the initial few infections are not of a singular or sporadic case in order to prevent unnecessary public panic if it were a false alarm. The length of the latency is complex, depending on the social–political structure of the city/national authority, usually from days to a week. While the authorities were waiting and monitoring the development of a potential outbreak from a small number, AI model that has been tapping on the vibe of the city beats round the clock, might have already sensed something unusual. For example, when a significant number of people started to show symptoms of an unknown disease, they would rant at the social media telling their circles of friends about their unwellness; those people and their friends and relatives would use search engines like Google, Bing and Baidu to seek information about the novel disease. This collective behaviour gives rise to surges of frequency of search keywords that could be picked up by web bots and Google Trend. Some web bots would be scrapping social blogs, tweets, opinions and comments posted on social media, harvesting hints of the illness-related sentiments and keywords, right from the patients and their peers.

Implementing this type of early warning prediction system requires a cooperative system of AI technologies. Its operation might have already been infiltrated into our daily lives without us knowing. The infiltration does not even intrude illegally into our privacy. Our modern society is accustomed to sharing and outreaching our private life in words, photos and videos, revealing our identities, background and even locations all the time. Similar to online advertising recommenders who collect logs of our online activities, these digital footprints are perfect ingredients to feed web bots, for them to understand trends and events of everybody's life. In the name of disease outbreak detection, the enabling AI technologies for hunting our information at the individual level include information retrieval [2], text mining [3] and NLP [4]. These three steps are conducted automatically running 24/7 across as many

social media and news platforms as possible, for absorbing information relevant to a disease outbreak in high quantity and quality. The steps form an end to end process, taking the collected information spied online on each individual social media user, analysed and infer about a predicted hypothesis—a collective signal that an unknown disease is among the people, and they are getting increasingly concerned about it. To take a step for validating this hypothesis further, macro-view data analytics is applied, the so-called big data analytics [5] which massively collect big data feeds from CCTV, IoT, mobile phone usage data, GPS, voices, emails, ATM transactions and any form of digital communication and activities transferred online [6]. Similar AI algorithms but different in designs and scalability for functioning in big data infrastructure are used. The analytic results from both macro- and micro-levels are augmented, in order to predict an outcome with higher certainty. In addition to the three steps in the prediction process, popular data analytics with AI are anomaly detection and correlation analysis and often with visualization. Anomaly detection, sometimes known as outlier detection or deviation detection is a method of recognizing abnormal events which happened along a timeline of normal events. Often the detected abnormal events are suspicious and warrant investigation. The detected rare events may lead to grave consequences if they are ignored. Picking out the events which are characterized by extraordinary values can be done either simply by computing the interquartile range of the variables or clustering which groups similar data points together, for data that are of two-dimensional (variable of changing values over time) and multi-dimensional (many of those variables), respectively. The rare events will reveal themselves as outliers (data points that stay far from the median) with respect to the majority of data points, either in the form of box-plots or visual clusters. In anomaly detection, outliers will stand beyond a scale that is centred by an Interquartile Range (IQR). Firstly, all the data are used to compute IQR which is comprised of 25, 50 and 75% of all the data. Finding outliers from IQR is very common in descriptive statistics. It is a measure of statistical dispersion, being equal to the difference between the upper and lower quartiles. Some assumptions would have to be made by the users, to find outliers from the data: outliers are those data points that fall below $Q1\text{-}threshold \times \text{IQR}$ or above $Q3 + threshold \times \text{IQR}$, where threshold typically takes a value of 1.5 which is called soft outlier threshold. It could be adjusted to be greater, called hard outlier threshold for identifying some very extreme outliers, e.g. at the value of 3 or larger (Fig. 3.1).

The IQR method is overly simple. The data points often are records of Tweets or sales patterns that would be well described by multiple attributes. In this case, clustering that makes use of Euclidean distance or Mahalanobis distance for measuring the non-linear similarity between data points is used. By the same concept as IQR, outliers are data points which fall outside of majority clusters. A visualization of outliers which lingering beyond the cluster that represents the main data distribution is shown in Fig. 3.2.

Correlation analysis is a statistical method which measures the closeness of the relationship between a pair of data series. It is useful for studying the relation between separate events which happen in pace at (almost) the same time or by (almost) the same trends. For predicting disease outbreak, often these two methods are used in

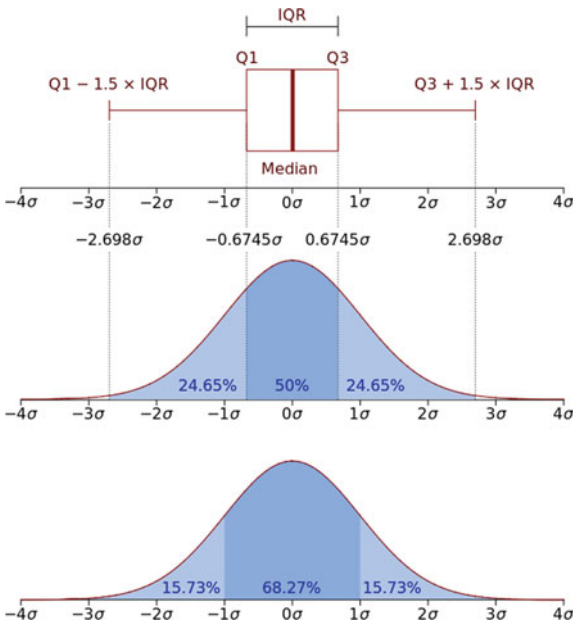

Fig. 3.1 Boxplot and a probability density function of a normal distribution of $N(0, \sigma 2)$ population

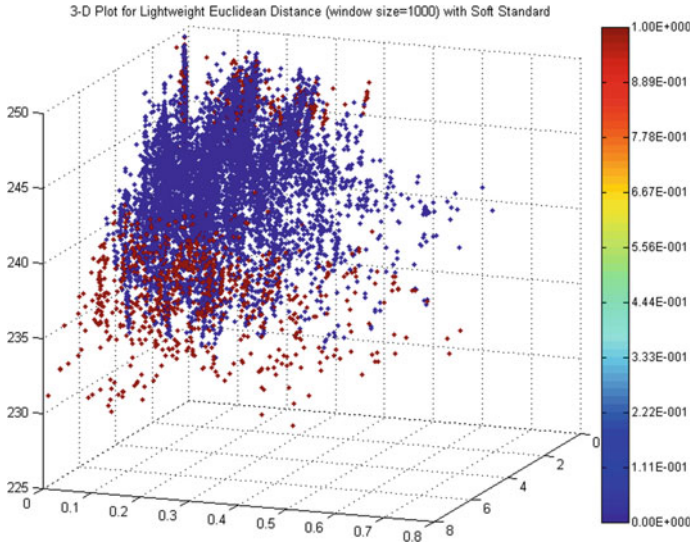

Fig. 3.2 Visualization of outliers detected by clustering using Euclidean Distance

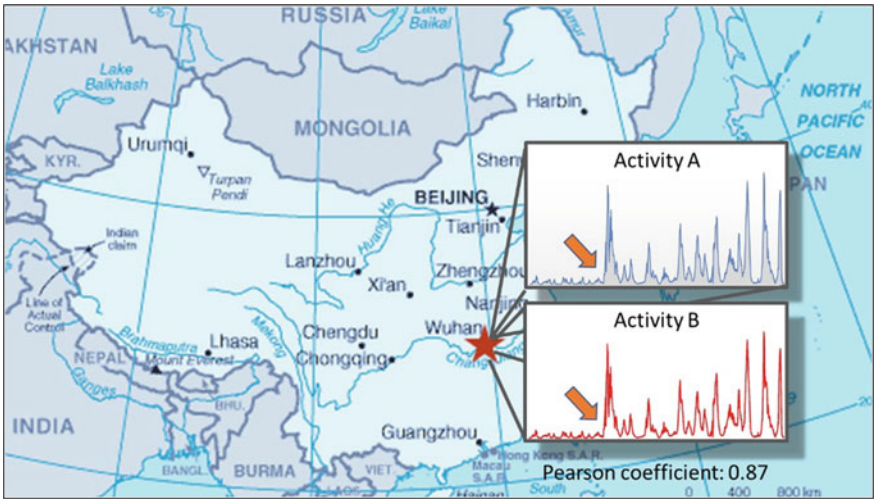

Fig. 3.3 Illustration of anomaly detection, correlation comparison and geo-map visualization

complementary to each other. For example, in a city it is detected that a surge of tweets complaining about their discomfort (or medical symptoms), at the same time people are panic-buying relevant symptoms-relief medicine from pharmacy stores. The rates of online posts and drug store purchases are correlated. Then it is pretty sure that some illness is going on in the population. These signals can be picked up by the prediction system and analysed using anomaly detection and correlation analysis. For decision support, the anomalies and their correlations could be visualized over a geographical map. An example is illustrated in Fig. 3.3 where two activities exhibited irregular patterns different from the past, and they arise at almost the same time same trends.

Being able to detect anomalies among the activities in a city triggers an alert. However, the prediction can be taken further from the moment and current situation to continuously predict about the future development of the outbreak. The latest ingredients from the data sources will continue to be tapped on, for further predicting how the outbreak will take place next, how many days before it will hit certain places, what the estimated number of infected cases will be and even how much damage it may cost. Further big data analytics could reference to the population of the nearby cities, the demographic details (percentage of vulnerable groups), the medical supports and intensity of social gatherings, hence the spreading rate, etc. These variables could be inputted to another prediction model which is formulated for extended prediction knowing that an outbreak is detected to be happening soon at a place. In this extended prediction process, there are a number of machine learning algorithms [7] that could be used, deep learning is one of the popular choices. This prediction process at the frontline requires the most up-to-date information which could come from the same data sources that were used to detect anomaly prior to the outbreak. Figure 3.4 shows a process diagram with two tiers of prediction modules.

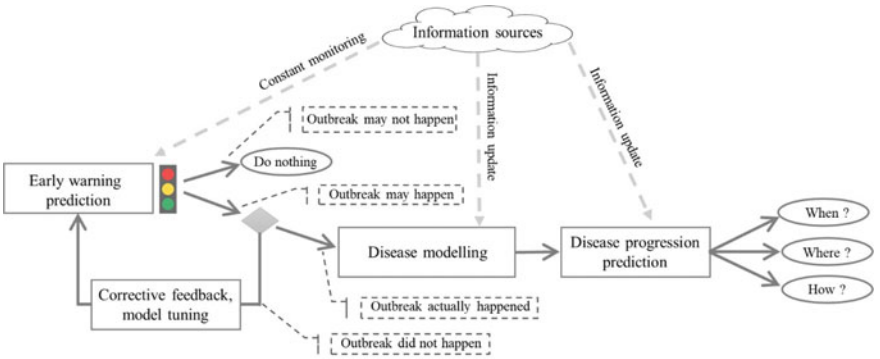

Fig. 3.4 A two-tier prediction model for predicting the next pandemic using big data

One is used for triggering an alert by continuous monitoring relevant activities. Once a potential outbreak is speculated to happen with a considerably high probability, the second prediction model will be launched to further analyse using the latest information available and predict a list of outcomes concerning the progress of the outbreak.

3.2 AI Predicts the Fate of COVID-19 Patients

Soon after the outbreak has begun, without exception, hospitals of every country are overwhelmed by many patients who are infected by the virus, causing huge stress on the medical teams. What if the likelihood of a COVID-19 patient developing into severe illness could be predicted? If the propensities of the clinical pathways for this novel disease could be predicted well, medical resources can be better allocated, managed and prepared in advance. Knowing the foreseen severe illness in advance means a more realistic estimation of demands in resources especially special equipment or those in tension which may require a long process for acquisition. The whole treatment process for every patient could be made more efficient.

Upon the urgency of needing an AI-supported clinical tool in predicting the fate of COVID-19 patients, a synergy is sparked between researchers from School of Medicine, New York University, USA and two Chinese hospitals, collaboratively develop a new AI prediction system [8]. During the project, they discovered that not all mild symptoms are equally important in turning a COVID-19 patient to severity of illness in a later stage. To most people's surprise, they observed that the popular symptoms such as high body temperature, early lung CT scans (before glass-like opacity is apparent), strength of body immunity and the patient's demographic details such as age, race, gender do not really predict well the consequence of serious disease. Testing out all possible variable factors, which are called predictors or attributes in data mining, over very limited number of patients at the beginning of the outbreak,

they found that all the popular mild symptoms do not predict well. However, three somewhat surprising factors do link strongly to the development of severe illness in the later stage. They are surge in haemoglobin levels found in the body, muscle aching (myalgia) and the subtle fluctuation of presence of enzyme called alanine aminotransferase in the liver. When these three predictors are used together, a relatively good accuracy level at 80% can be achieved in predicting one of the later-stage illness known as Acute Respiratory Distress Syndrome (ARDS). In layman's term, ARDS is the difficulty of breathing which took many lives of COVID-19 patients at the end.

How this prediction can be done, which can attain up to 80% accuracy over a small available sample set using only three predictors? In AI, this is predictive analytics, which is about training up a representative model that recognizes the relations or mapping between the attributes of the dataset and the predicted outcome, which is simply binary—Yes, it will lead to ARDS or otherwise. The underlying relations at the predictive model could be highly non-linear considering that a number of attributes can take on very different values, out of many possible combinations, there is one that forms a strong decision path that points to the predicted outcome. The rest of the combinations and their links rather look random and meaningless (no strong link exists).

The predictive model would consist of two round of selections—feature selection as pre-processing step and random forest which selects the best performing decision tree from many other optional trees. Feature selection is basically a feature engineering process which transforms a full set of features with, however, maximum number of features (attributes) that describe the data to only a feature subset, that is, significant enough for making a reasonably accurate prediction. In the case of ARDS, the subset contains only three features, the liver enzyme, myalgia and haemoglobin level, that are sufficient to make an accurate prediction in lieu of the whole feature set which might contain other possible symptoms and patient's demographic details. It is known that given the number of maximum features is m, there is 2^m number of possible candidate feature subsets. For $m = 64$, the number of candidates is $2^{64} = 18446744073709551616$ (20 digits) which is an astronomical number. It implies that the search must cycle through that many times of repeating the tests over all the possible candidate subset by brute-force. An alternative search is by metaheuristic search or swarm search [9]. Several randomly chosen features are picked into a candidate subset at the beginning; using heuristics the candidate subset improves its selection of features in each round. The feature subset is refined in each round, until a predefined maximum number of iterations is reached or no more significant gain in the marginal improvement between successive cycles is observed. At each iteration, a slightly modified feature subset is put into the predictive model building process, for testing out the goodness of the current feature subset. The flowchart of swarm feature selection is shown in Fig. 3.5.

The iteration time varies according to the workflow depicted in Fig. 3.5. However, what most consuming probably is the core of the selection operation, namely, the 'classification model training' module. It basically builds a predictive model based on the input candidate feature subset each time. Each time after a predictive model is

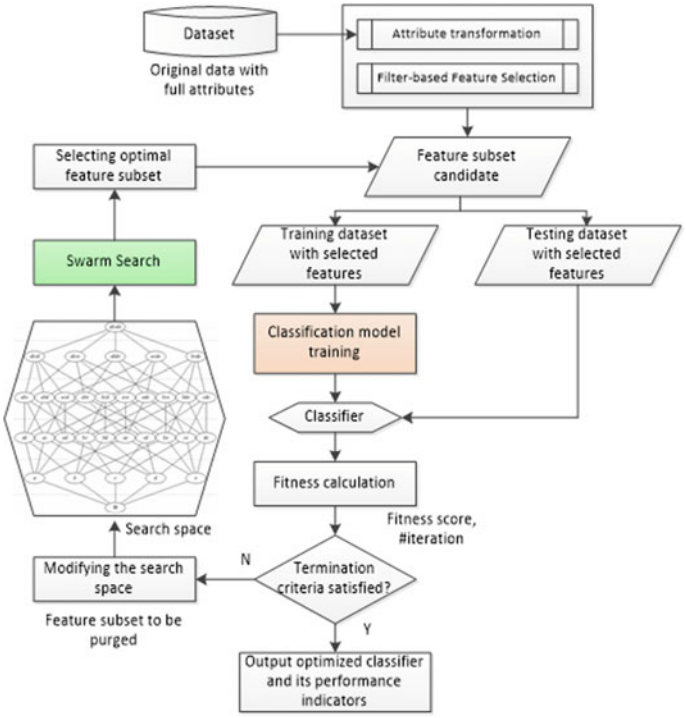

Fig. 3.5 Flowchart of swarm search-based feature selection method

built, it is put under test using the available testing data; the performance of the model is recorded as a fitness value which could be the average error in the prediction test using the candidate feature subset.Certain computing power requirement imposed on the hardware is required for fulfilling this iterative testing.

On the other hand, a Random Forest (RF) [10] which is an ensemble of decision trees in tournament is needed for assuring the highest possible level of accuracy in the final predictive model. RF is a competitive algorithm specialized in finding the best form of decision tree in doing classification task which tells apart the patients that will develop into ARDS from those otherwise. RF tries various combinations of model configuration parameters, different portions of training samples and even different features (which is unnecessary here since we have swarm search feature selection as pre-processing). It tries to exploit the memory space of the computer in trying out different individual decision trees which are dissimilar to each other, forming a collection of decision trees made up of different configurations. Then RF conducts a committee voting to vote for one that is more accurate than the rest of other individual trees, nominating that tree as a final winner. The winning tree is the best of all, taken as the final decision tree model at the end. In the context of using AI to predict ARDS consequence for COVID-19 patients, the final winning decision tree is used as decision support. Each branch of the final winning decision tree is traversed

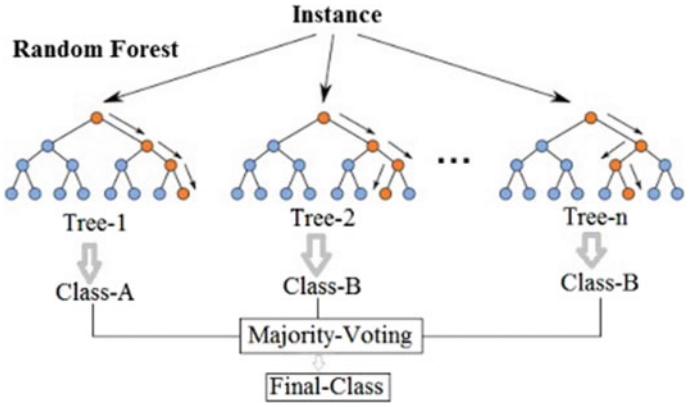

Fig. 3.6 Illustration of generating different optional trees in Random Forest

by testing each option at each tree nodes along the tree path, eventually after a series of chosen options, it leads to a conclusion at the leave (bottom level of the tree) indicating whether or not the patient will be suffering from ARDS. An illustration of RF is shown in Fig. 3.6 where each optional tree is built by using different portions of training samples, in order to generate a pool of different trees for performance selection. In this project, however, the training sample size is limited, instead different parameter values and configurations are tried, together with the swarm search feature selection, in finding out the best decision tree as the end output.

3.3 Finding the Most Accurate Predictive Analytics

Trends of disease spread can be monitored by data analytics and statistics. Predictive analytics which is a major branch of AI can forecast by regression and identify by classification where and how much healthcare demands are anticipated. So, resource planning, acquisition and allocation can be facilitated more precisely and timely. By knowing the spread patterns and fusing the trends with consideration of other social factors, intervention can be better applied. However, predictive analytics is not guaranteed to deliver perfect prediction every time, although it is known to do better than human educated guesses. Rarely a100% accuracy can always be achieved.

A tech giant Alibaba developed a one-stop AI system [11] solution at the urgency of relieving the overloads from the frontline healthcare workers. The AI system claims that an accuracy of 96% can be achieved in distinguishing between coronavirus infection in CT scans of patients from other pneumonia cases. The advantage of the AI system is more on the speed and efficiency which takes typically 20 s for a decision rather than a quarter of an hour when done by human experts which are heavily overloaded. The AI model is trained with 5000 empirical CT scans from confirmed coronavirus patients collected during the initial period of outbreak.

US military announced an AI algorithm [12], that is, able to predict infection 2 days ahead of clinical suspicion, before apparent symptoms can be confirmed, at accuracy level of 85%. The model is trained by using more forty thousand cases which are collected globally and 165 biomarkers. Early warning of a person who is likely to develop into COVID-19 patient, made possible by wearing wearable sensors that are strapped over the chest and wrists for biosignal monitoring. From the complex relations between the vital signal streams from each biomarker, the AI algorithm will know if the person will fall ill in the next 48 h before symptoms emerge. Again, in this technology, time is a priority for early warning and saving the potential risk of identifying a virus carrier that may infect other crew members if he is not detected early.

It can be seen that early prediction followed by detection, confirmation and intervention from Fig. 3.7 is essential to enhance the medical treatment process from one end to another. Predictive analytics plays an important role, despite speed and efficiency, accuracy is important. With the current glooming figures of over 2 millions infected cases and over 160 thousands deaths, even a slight percentage of errors from predictive analytics means putting thousands of lives at risk of false-negative detection. Section 3.2 discussed optimized feature selection and ensemble decision trees help sharpening the accuracy of a prediction model. However, these fine-tuning techniques are applicable to predictive analytics at individual level, i.e. predicting the outcome of a particular patient. The prediction is usually done by training–testing phases of repetition until a satisfactory level of accuracy could be reached; or else, the input variables, such as selection of training data, choice of method of feature selection, choice of predictive algorithm and its parameter values, should be re-tuned.

Timeline of COVID-19 disease progression		Predictive surveillance	Clinical surveillance
Infection	Hit by the virus, without even knowing	Early warning before symptoms appear	
Feeling unwell	Body starts to weaken		
Symptoms	Apparent symptoms arise, such as fever, body aching and dry cough	Prediction the likilihood of infection	
Home medication	Pain killer or symptom relief medicine (in the hope that mild symptoms may go away)	Oprional: Telemedicine remote consultation	
Hospital admission	Needing medical attendance		
Screening	Basic observation at triage, to determine the appropriate direction of treatment	Classify the patients into group (for decision support only)	Basic body check
Infection testing	Throat Swab Culture, Rapid Influenza Antigen, Blood Test, CT Scan, etc		Infection identified and confirmed
Treatment	Medication and other medical intervention, including respiratory aid and ICU if necessary	Predictive pathway (for decision support only)	Traditional clinicial pathways
Discharge	Leaving hospital because the patient is recovered or deceased	Accumulate medical history as record, for future predictive model improvement	

Time

Fig. 3.7 Timeliness of AI predictive analytics in disease progression

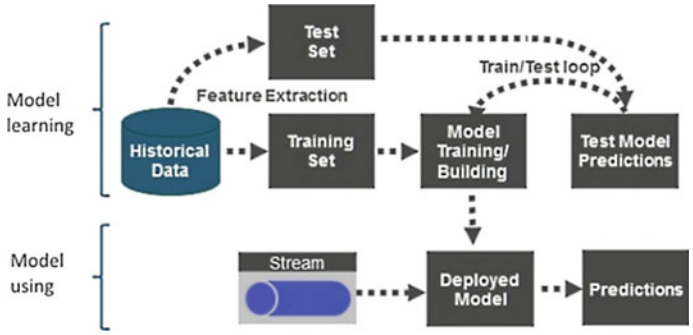

Fig. 3.8 The traditional training and testing process in building a predictive model from data feed for medical prediction

When a model is trained sufficiently to maturity, new data feed is subject to the model as unseen testing data to the model for making a prediction.

The train-and-test process shown in Fig. 3.8 is generic, suitable for predictive analytics for both individual patient and the trends of disease spread for a population. The advantage is the ability to observe the performance output at the end of the model training before it could be put into actual use. Knowing the performance and scoring them quantitively provides an opportunity for fine-tuning the prediction model. Being able to benchmark a prediction model is very important in this case, because we have several layers of uncertainty: not all algorithms generate good results; the model performance is sensitive to the training and testing data, sensitive to parameters and configuration of the model and features that characterize the data. Tuning up the best performing model is a tricky task because multiple health signals and variation in training records. This problem is worse especially in the beginning of the outbreak where the available samples are few and variation is high. But ironically a reliable model is more demanded for rushing out to the deployment at this crucial time.

The underlying reason for the difficulty in tuning up an accurate model is the fact that a set of heterogenous data are mixed and inputted as one training set to the model construction. These data come from logs of medical procedures, drugs and antibiotic administered, their responses from the body reaction, and the body conditions reflected from the vital signals. The interactions between these variables and their effects are complex, bound to be naturally non-linear in their relationship with respective to the predicted outcomes.

A methodology for finding the best AI algorithm is recently published, called Group of Optimized and Multi-source Selection (GROOMS) [13]. GROOMS is created with an objective of finding a model that gives the best prediction accuracy under the shortcoming of having limited available and little knowledge about the novel disease. GROOMS allows ensemble forecasting similar to random forest, by generating a group of forecasting models with parameter values tuned; some candidate models can take on several input data sources since medical forecasting often relies on multiple sources as mentioned above.

As shown in Fig. 3.9 GROOMS methodology allows loading in a small dataset sample which may be all that is available in the early time; passing them through three passages from top to bottom. A number of candidate models are tested on the data, each of which has its parameter values tuned to optimal. Some models are multiple regression which embrace multi-modal data from different sources. There are generally three types of forecasting algorithms in the AI family: 1) non-parametric models that have no parameter except a single input variable from a time-series input. 2) parametric models that have multiple parameters that are sensitive to performance and need to be tuned for the best performance. For instance, decision trees have parameters about splitting criteria and minimum or maximum number of nodes/depths of tree, etc. Neural networks have parameters about learning rate, momentum rate, network structure activation function, etc. 3) dual models—machine learning models that are sensitive to both of data input and the model parameters; they generally require most time and most difficult to tune.

The candidate models from the three passages are constructed, tuned, tested and performance is scored. This may repeat, however, times it takes, until each candidate model gets tuned up to its best performance. When they are ready, all the candidate models are subject to a committee voting assessment like a panel section. For regression style prediction models, the average fitting errors, e.g. RMSE is taken as a performance score. The panel selects a winner model which can produce a forecast

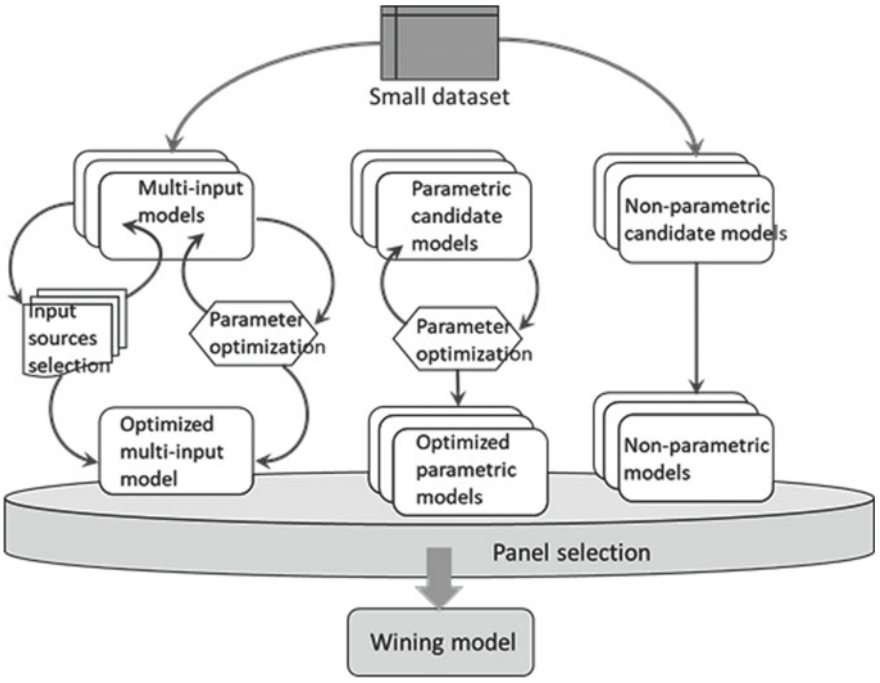

Fig. 3.9 GROOMS methodology

Fig. 3.10 GROOMS in action, evaluating AI algorithms in order from simple to complex

or prediction at the lowest error rate. The forecast or predicted outcome is therefore deemed to be one that is best available from what is available on hand.

How do data scientists go about using GROOMS given there are various sorts of machine learning and data analytics available in the big family of AI? The techniques differ in complexity, speed in modelling and sensitivity to parameter tuning, which can be roughly placed in five groups. The order of evaluating all the techniques is advocated as the sequence shown in Fig. 3.10. If time is not of constrained, all the techniques could be tried for selecting the best which may take days or weeks depending on the computing hardware available and the volume and dimension of the data. If time is urgent, especially during the critical moments of rescues when some scientific analysis is needed urgently, it is advised to progressively tested from simple techniques to sophisticated but time-consuming techniques.

The techniques are ranked in complexity in five groups, where Group 5 is the simplest, and Group 1 is the most complex. Although the exact time taken for each technique in each group is quite unpredictable, it is generally known that computation involves iteratively refinement and complex matrix operation would take much longer than basic statistical inference, e.g. moving average, which works over the data in just one pass. The five groups of techniques generally are as follows.

Group 05: Time-series forecasting by econometrics principles—this group of data analytics attempts to find a curve-fitting curve over the actual data curve, interpolating the fitted curve to the future horizon for estimating forecast. Popular choices are regression, auto-/exponential regression, integrative autoregression moving average, etc. Although this group is not the simplest among the five, they are fundamental and should attempted first as classical methods. The results are trustworthy in view of many statisticians.

Group 04: Basic data analytics—this group serves as a complementary role supporting the results of group 1 techniques and results. They expand the description of the data on the statistics landscape, by charting the statistics results. Common techniques are descriptive statistics, where the mean, median, min/max, variance and standard deviation are computed, frequency distribution, histograms, scatter-plots, box-plots, etc., e.g. average, min, max, the increments since yesterday or some days ago, rate of increases of suspected cases in comparison to cured cases, etc. These visualization charts, though basic, are some of the most popular techniques that most countries are using now to monitor and track the development of the virus spread. Government concern about the accelerating rate which is shown as the gradient of a curve of confirmed/suspected cases during the outbreak. The so-called flattening the curve means the gradient as daily increase is kept below certain threshold, so that the hospitals and medical infrastructure remain capable to treat new patients. It is alarming if ever the growth of the curve exceeds the threshold, implying that the medical system is collapsing no longer being able to handle new patients. The consequence is disastrous, as infected cannot be treated, they could only come home after refused by hospital, continue to infect their families and neighbours like a chain reaction. The threshold is a user-defined arbitrary variable that should be set according to the medical capacity of a country which is different country to country. The general rule of thumb is to keep the curve well under the threshold as much as possible imposing lockdown measure, discouraging and even forbidden anybody to travel out unnecessarily. Some of these typical charts are shown on COVID-19 dashboard, programmed in Tableau which is an interactive visualization software, in Fig. 3.11.

Group 03: Lightweight machine learning algorithms for forecasting—simple machine learning algorithms like those called 'lazy learners' or 'incremental learners' are used as base learners in forecasting. In contrast to conventional machine learning algorithms based on greedy-search, the incremental leaners learn to approximate the fitting curve to the actual curve, by updating the model on pass of data at a time. Incremental learners run fast on par with their counterparts, producing reasonably accuracy performance. Relatively they are also prone to misfitting by either overfitting or underfitting the training data. Tuning up the parameter properly is necessary for building a useful model.

Group 02: Complex machine learning algorithms for forecasting—complex machine learning algorithms are used as base-learner for doing forecasting. These algorithms in general are capable of recognizing and handling the non-linearity from complex data series that are twisted with jitters and large magnitudes of fluctuation. The model requires intensive parameter tuning too because the accuracy performance is very sensitive to the model configuration. For examples, SVM takes several important parameters each of which will impact greatly on the accuracy performance as well as the final outcome. Depending on the implementation of the algorithm, typical parameters are regularization variable of error term, kernel (poly, Sigmoid, rbf, or linear), degree of the kernel function for the polynomial (if the kernel = poly), gamma and initiation state, etc.

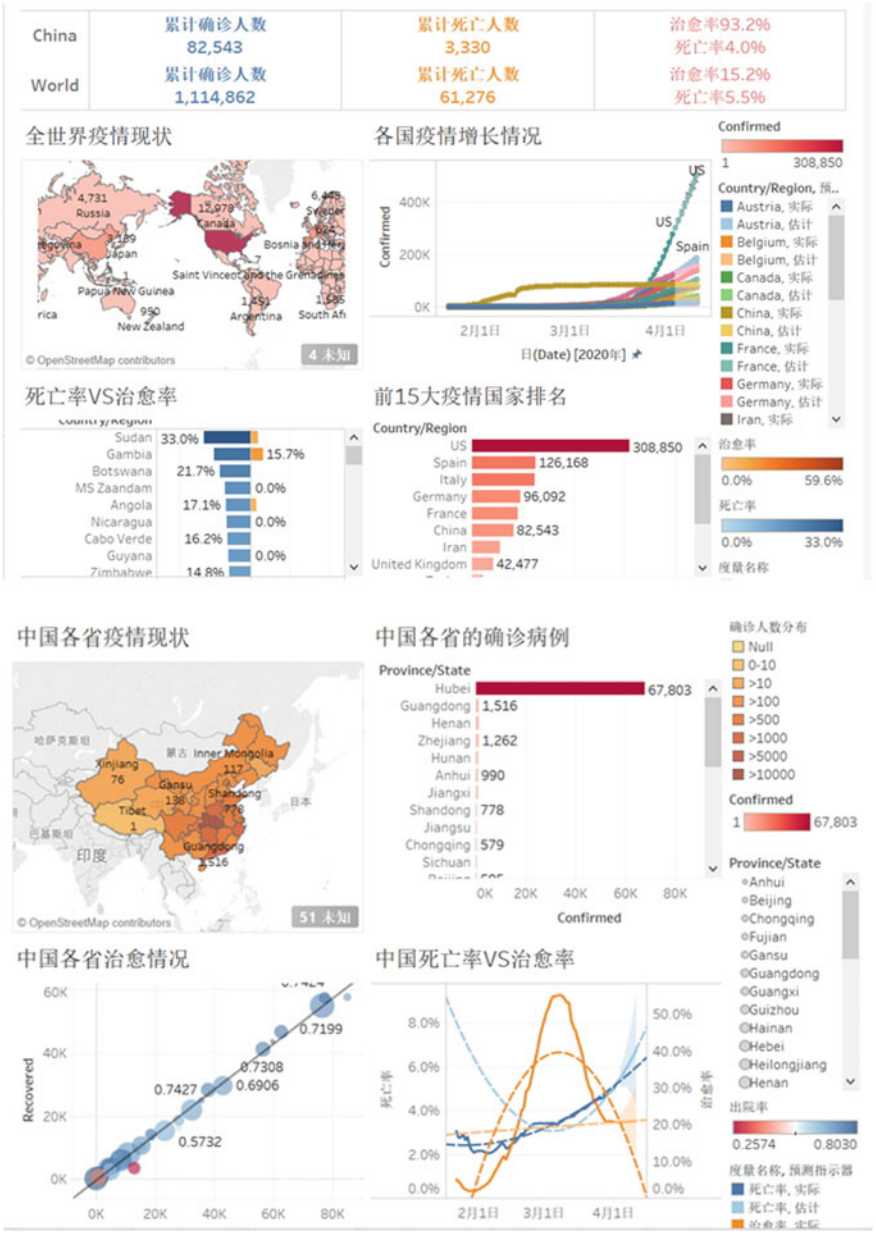

Fig. 3.11 Examples of COVID-19 dashboards using simple charts programmed in Tableau

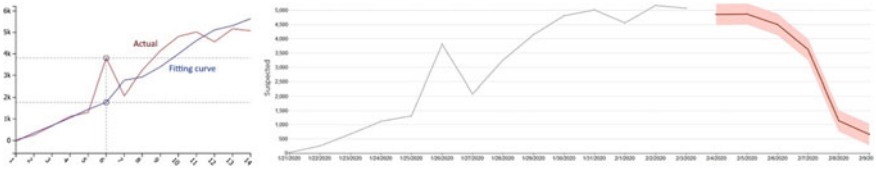

Fig. 3.12 Comparison of early forecasts by simple predictive analytics and the complex algorithm. Left: Holt-Winter, Right: Polynomial Neural Network

Group 01: Complex machine learning algorithms with multiple regressions for forecasting—very non-linear machine learning model, such as neural network that takes many options of how the network structure can be configured and the associating parameters and activation function. Deep learning belongs to this category which has even more options in convolution feature maps set up in addition to the default neural network setup. This category is suitable for small and limited data but correlated data from multiple sources. For prediction of COVID-19 virus spread, forecasting in early stage is important; but data is scarce in short or very little history. Multiple regression may be used here, feeding in a collection of related data as inputs. Model tune-up involves the parameters for the model itself and the pre-processing of a group of time-series.

Figure 3.12 shows a comparative forecasting example over the early time-series of daily increase of confirmed COVID-19 cases. The forecasts are made by Holt-Winter algorithm which is a popular econometric from Group 5 and Polynomial Neural Network with multiple regression from Group 1. It can be seen that the forecasts by the two groups of methods lead to totally different outcomes. The Holt-Winter method from Group 1 placed too much empathizes on the trend of the data curve, the forecast is almost a straight line that follows the same gradient from the past records. In contrast, the Polynomial Neural Network from Group 5 that inputs the other relevant data series offers a reasonable forecast curve with a gradual downward trend and few small turns. At the end of writing, it is known that the daily increase cases for China have dropped to a single-digit figure close to none. The Group 5 algorithm predicted so several weeks ago prior to the actual drop that really had happened. In summary, one must be careful in choosing the right AI algorithm in prediction especially if the prediction involves life and death. If it is urgent, lightweight approaches could be used. If time is allowed, it worth investing the time spent on tuning up the parameters as well as considering other relevant data series together, for a quality forecasting model,

3.4 Predicting the Virus Spread by SIR and SEIR Models

In epidemiology, mathematical modelling is used to estimate the spread of the disease considering important and dynamic variables which are inter-depending on one

another. Unlike time-series forecasting which is purely based on historical records from which a forecast is projected to the future horizon, mathematical modelling based on compartment model is used involving three or more inter-link factors. The accuracy of prediction from time-series forecasting assumes the future situations remain unchanged, which is unrealistic. The forecast is merely a projection of what happened in the past, following up the same trend from the fitted curve which is again inferred based on history.

The mathematical modelling that is favoured by epidemiologists is inherited from the concept of compartment model. Compartment model has been used to simulate transport of quantified items from one state to another, often within a closed-loop and bounded space. For example, moving one kilogram of ice from a freezer room to a kitchen at room temperature, you will have 1 kg of less ice from the freezer room but increase by 1.091 litre of water, assuming it is a closed system without evaporation, leaks or vents through which the material will escape to elsewhere. A simulation can be set up thereby the room temperature in kitchen could be tested as a variable against the amount of water as liquid that exists in the kitchen compartment. By the same analogy, a simulation can be set up to study the inter-relationships between the three major variables of which their existences are interlinking and proportional. The three common variables in epidemiology are Susceptibles (S)—how many people are healthy but susceptible to virus infection, Infected (I)—how many people who were healthy and now become infected and Recovered (R)—how many people who were infected now have recovered. They are equivalent within a confined system. It is noted that the transitions between these three variables are unidirectional (S $\rightarrow$ I $\rightarrow$ R) and the relations are proportional. The more people are infected the less healthy ones we got, infested and recovered numbers are somewhat correlated. Taking time as a system variable, as days pass by, one can observe how many people remain susceptible, infested and recovered. At any point of time, the values of the three variables can be projected and observable along the horizon of future days. For example, given a fixed population, we will know the numbers for S, I and R, and their interaction as time progresses, when we define some changing rates from S to I and I to R. The rates are known as contagious rate (or reproduction rate) and recovering rate, respectively.

The S.I.R. model was theorized as a mathematical model by Kermack and McKendrick in 1927 [14], coined after the definitions of the three compartments S, I and R. SIR model is one of the most fundamental models, based on which other variants are developed to model epidemics and pandemics of different characteristics [15]. Taking time t (in days) as a natural and discrete variable which will pass day after day regardless of whatever will happen on earth, the dependent variables in the model are defined as a function of time:

- $S(t)$ = number of susceptible persons at time t. Everybody is assumed to be born susceptible to virus infection. For COVID-19, it is assumed that susceptible (healthy) individuals who were infected and once they managed to recover, will be immuned. So everybody can only be infected once in their lifetime in this

scenario although there are some rare cases as exception that a same person gets infested twice or more.

- $I(t)$ = number of infected persons at time t. These are the people who had contracted the disease and able to pass on the disease to susceptible people, regardless of symptomatic or asymptomatic. Once infected, the individual will only move on to the R stage, as a matter of time. I_0 is the zero patient, $I(0)$ is originally how many of such zero patient exists at the beginning. Usually $I(0)$ is unknown or by default set to 1.
- $R(t)$ = the number of recovered persons at time t. R patients are those came from I but either recovered or died at the end. In either way, R patients will not be infected again, neither can they become S nor I anymore.

As such, converting the functions in number above to percentage, we have $s(t) = \frac{S(t)}{N}$, $i(t) = \frac{I(t)}{N}$, and $r(t) = \frac{R(t)}{N}$, where N is the number of population which remains unchanged throughout the simulation. Births and migration within the simulation period are not considered in the model. Then summing up the percentages $s(t) + i(t) + r(t) = 1$ in the close-loop system, where $s(t)$ and $i(t)$ are directly inverse proportional to each other, $i(t)$ and $r(t)$ are associated with a decay function. The model requires additional assumptions to maintain the validity. $S(t)$ will only decrease or remain the same number since an outbreak started, the rate of decline is subject to the number of susceptibles who can remain isolated, the number infested who can infest, and the intensity of contact level between the remaining susceptibles and the increasing number of infested. Let β be the number of contacts through which each infected person will infect susceptibles per day, assuming that the virus is absolutely contagious. That means once any single disease carrier contacted the previously unexposed susceptible, he will be infected for sure. β is known as a reproduction rate by which the disease spreads. At the beginning, everybody is susceptible to COVID-19 (assuming nobody is born immune). If β is equal to 1, each infested will pass the disease on to a susceptible. If β is smaller than 1, the outbreak will not sustain because the number of people to be infected will be less than the number of existing infested in the future; the pandemic will die out sooner or later. But if β is greater than 1, the population will experience an exponential growth of infection.

By referring to past experiences of similar respiratory pandemics, the reproduction rate β for COVID-19 is estimated to be around 1.5 and in the worst scenario β can be up to 4. It was conservatively calibrated to 2 in the initial two months of outbreak since mid-December 2019. By $\beta = 2$, the initial infested person would spread the disease to two others, each of these two others will continue to spread to another two others, and so forth the disease propagates. The growth rate of the disease spread is drastically as sharp as exponential, but it is not uncommon in the early stage of outbreak. However, assume no control is applied to stop or slow down the spread, more than a thousand people could be infected starting from only one person after 10 days. Given another 10 days down the road, the total number of infected people be over a million. And this will continue to multiply to tens or hundreds of millions in a worst-case scenario in the near future. Such sharp rises can be seen in some countries in terms of death tolls in Fig. 3.13. The death tolls are very much proportional to the

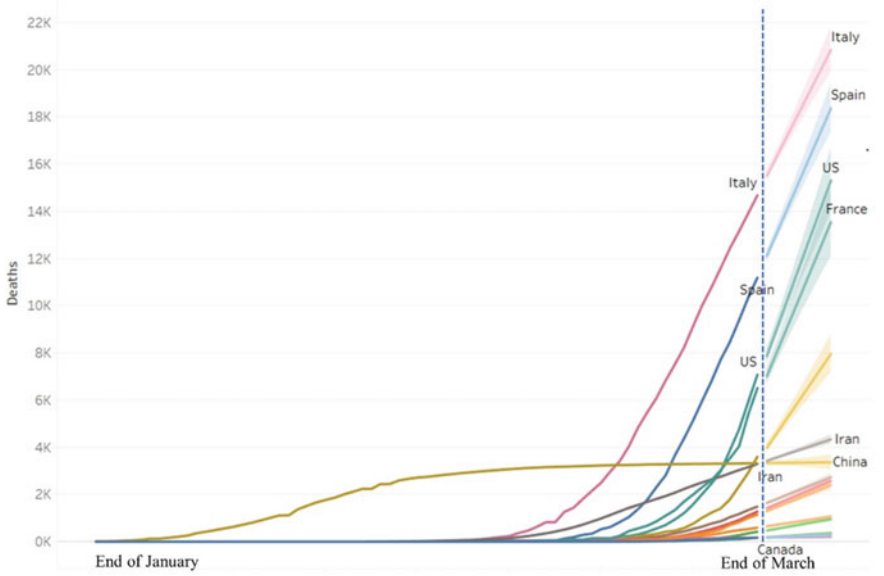

Fig. 3.13 Sharp rises of death tolls in US and European countries at the beginning of COVID-19 (data source: https://www.tableau.com/covid-19-coronavirus-data-resources)

number of infected patients. It is known about 5% in average in the world; however, this varies from country to country. Figure 3.14 shows a tornado chart of the top highest Case Fatality Rate (CFR) as of end of March 2020. Although there are many factors that lead to death, such as age, strength of body immunity and pre-existing illness conditions, the increase of number of infected people has a direct impact on the capacity of hospital upon which the chances of a patient's recovery or otherwise rely on. To the end of this, keeping the number of infected patients under control while protecting the susceptibles is a prime objective, and S.I.R model is going to inform us about the future scenarios of how far we are from these two equivalent targets.

To simulate the dynamic of the S.I.R. model, we let the transition probability from compartments S to I as $\frac{\beta}{N} \times \omega$ where ω is infective index (having $\omega = 1$ means for certain each contact will get a susceptible infected), and from compartments I to R, the transition probability is governed by γ which is mortality rate or removal rate. γ is the average number of infected patients who recovered or died per day over the total number of people who are currently infected on the same day. The dynamics of the S.I.R. system mainly depend on two forces—how contagious that a susceptible became infected and how soon an infected individual can be removed from the system. The removal rate is directly hinged on how much capacity a medical system can provide in quickly testing, identifying and confirming the suspected individual, and perhaps quarantine them in time before he goes around and infects other people. This issue relates to the availability of test-kits and the efficiency of the

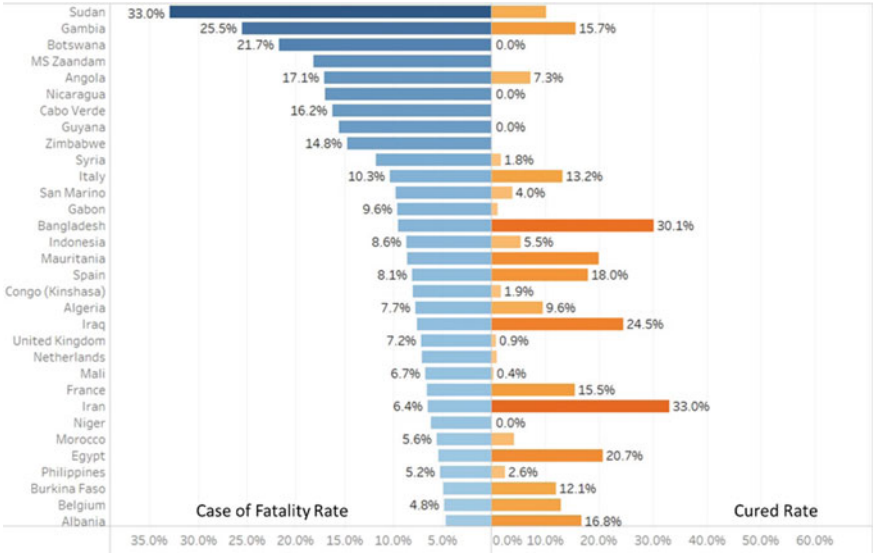

Fig. 3.14 Top-ranked countries with highest CFR in COVID-19 pandemic (*Data source* https://www.tableau.com/covid-19-coronavirus-data-resources)

medical team, as well as the demands for these resources depending on the number of infected people. In the previous chapters, AI and technologies have been discussed in improving the medical capacity and efficiency. Nevertheless, the medical capacity assuming is a precious resource which shall not be overloaded even we have the best AI and best medical team. Keeping β as low as possible is imperative here. On average, $\beta \times s(t)$ number of new infected people are generated every day. These newly increasing infested people who supposedly need medical attention and the additional supply of medical resources are sitting on a seesaw; unfortunately, medical resources cannot be easily increased in a short time, but the newly infected people pile up in thousands every day in some countries. To make matter worse, infected patients who are rejected from hospitals due to over-capacity may roam around or at home infecting family members or other caregivers.

The S.I.R. model is constituted by the kinetic energies of three equations—the Susceptible equation, the Recovered equation and the Infected Equation in Eqs. (3.1), (3.2) and (3.3) as below:

$$\frac{ds}{dt} = -\beta \times s(t) \times i(t) \tag{3.1}$$

$$\frac{dr}{dt} = \gamma \times i(t) \tag{3.2}$$

$$\frac{di}{dt} = \beta \times s(t) \times i(t) - \gamma \times i(t) \tag{3.3}$$

$$\frac{ds}{dt} + \frac{di}{dt} + \frac{dr}{dt} = 0 \tag{3.4}$$

Equation (3.4) shows that the S.I.R. would have to remain balanced as an equivalent closed-loop system. So, for an example of setting up a S.I.R. model for COVID-19 outbreak that started on the first day in Wuhan, the city has a population of 11.08 million; assume there were 10 people who were infected and nobody had died of or recovered from the disease yet on the first day of outbreak. $S(0) = 11{,}080{,}000$; $I(0) = 0$; $R(0) = 0$. The complete S.I.R. model would therefore be

$$\begin{array}{rcl} \frac{ds}{dt} = -\beta \times s(t) \times i(t) & \leftarrow & s(0) = 1 \\ \frac{di}{dt} = \beta \times s(t) \times i(t) - \gamma \times i(t) & \leftarrow & i(0) = 9.03 \times 10^{-7} \\ \frac{dr}{dt} = \gamma \times i(t) & \leftarrow & r(0) = 0 \end{array} \tag{3.5}$$

In the early days of the outbreak, the coronavirus was novel. It is an educated guess initially to assume the values for the parameters β and γ to start the model. Then as the spread progresses and more information is collected, the values for these two variables would be adjusted to fit the actual scenarios. For an example and sake of illustration of how S.I.R. model works, the following assumptions are made: a small city with population of 100,000; an infected patient has an average timespan of 25 days of being infectious until either he recovered or passed away, so $\gamma = 0.04$; everyday every infested carrier will infest between 2 and 5 susceptible people; therefore, a varying β is adjustable in the S.I.R. model such that $\beta = 1$, 2, 3, 4 and 5. Figure 3.15 shows four possible outcomes generated by the S.I.R. model when β varies from 1 to 5.

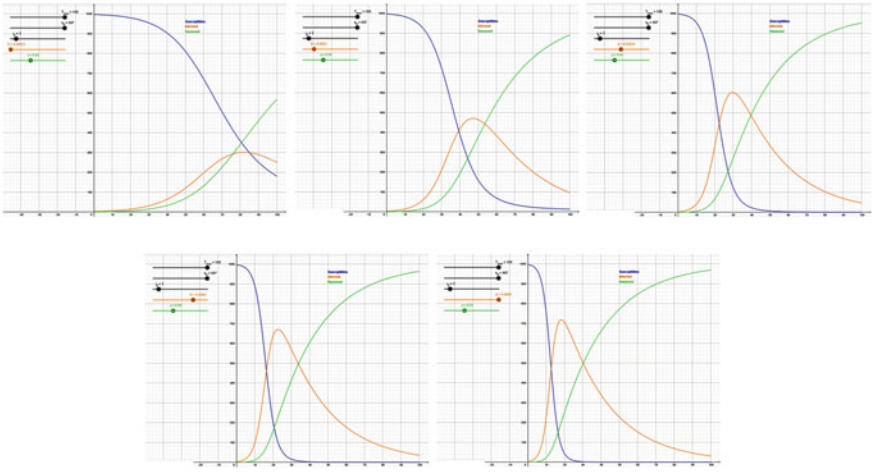

Fig. 3.15 Predicted scenarios by S.I.R. model with varying $\beta = 1$, 2, 3, 4 and 5 per population per day

It is noted that the output of the S.I.R. has three curves representing the future forecasts of the number of susceptibles, infested and recovered, respectively, on the horizon of future days. Mainly, these three curves depict the developments of the virus outbreak in different level of human contacts by β, starting with a small number of infested people escalates to a large number in a length of time. The red curves which represent the number of infested people are rising at a high rate, the more number of contacts per carrier per day, the sharper the gradient of the trajectory is. In all the four cases of simulation results, the trajectories will peak at different heights but sooner or later the curves slope down gradually due to the running out of fresh susceptibles to be infected and the recovery of those that were infected earlier. It is obvious that when the contact level of the population is set to minimum, the area under the susceptibles curve remains large. That is, the situation that every government wants to achieve, by keep most of the population spare from being infected. The other important point to be observed from the S.I.R. output is that the bell curves of the infected which may not be perfectly symmetrical should be kept as flat as possible. Or else it will exceed certain threshold where the hospital capacity is at maximum, once the curve goes beyond the threshold the hospital starts to get overloaded. If this situation is not remedied, the medical infrastructure will collapse and the frontline nurses and doctors in hospital get infected too as protective supplies run out and overworked to fatigue.

At all costs, the S.I.R. is telling us it is extremely mandatory to keep β at bay, preferably down to zero if possible though it means massive shut down in towns and cities, economy will be damaged. Assuming new vaccine will not be available any time in near future, social lockdown by social distancing, which is proven to be an effective measure, is necessary to keep a majority of population safe and wait for the virus to die out. Therefore, an important component in the S.I.R. model is missing, and the model needs to be extended to accommodate a compartment for the incubation period as 14 days is typical for COVID-19 and social distancing. Social distancing not only helps to keep β low by cutting down non-essential human interaction, but it also provides a space and time buffer for the person who was infected to seek medical help before he becomes infectious by his symptoms such as coughing. When social distancing is enforced, the just infested person as soon as he feels unwell, he would call CDC or any COVID-19 hotline. Thereafter, medical team would arrive his house, test him and escort him to hospital for quarantine during his home-safe period under national lockdown. This would result in early treatment and higher chance of successful cure. Furthermore, this latency provides opportunity for AI to help, such as telemedicine, timely detection as suspected cases and optimized hospital resource preparation.

To embrace this incubation or onset period of COVID-19 infection, S.I.R. model is, hence, modified to S.E.I.R model [16] which is suitable for inclusion of a latency period for COVID-19 between the compartments of Susceptible to Infected. Mathematically, for S.E.I.R. model, let $E(t)$ be the number of people per day exposed to the virus, infected but not infectious yet. $e(t)$ is the corresponding ratio of $E(t)$ per day per population. The mean latent period and the mean infectious period for COVID-19 are assumed to be $\frac{1}{\alpha}$ and $\frac{1}{\gamma}$ respectively, β is a rate that changes in time as

$\beta(t)$ where β_0 is the initial β value, and μ is the birth and death rates which are equal. The same assumption which states that no vaccine is available, and the recovered patients are permanently immune apply here. The additional Exposed equation and the other modified equations are defined as follow:

$$\frac{ds}{dt} = \mu - \beta(t) \times s(t) - \mu \times s(t) \tag{3.6}$$

$$\frac{de}{dt} = \beta(t) \times s(t) \times i(t) - (\mu + \alpha) \times e(t) \tag{3.7}$$

$$\frac{di}{dt} = \alpha \times e(t) - (\mu + \gamma) \times i(t) \tag{3.8}$$

$$\frac{ds}{dt} + \frac{de}{dt} + \frac{di}{dt} + \frac{dr}{dt} = 0 \tag{3.9}$$

$$\frac{\alpha \times \beta_0}{(\mu + \gamma)(\mu + \alpha)} = R_0 \tag{3.10}$$

R_0 is a significant indicator informing us how contagious the virus is [17]. It is known as Reproduction rate in S.E.I.R. model, defined by the mean number of susceptibles that an infected person will spread the virus to per day. So far, the exact reproduction rate for COVID-19 is not known yet, but it is estimated from the existing data from WHO to be ranging from 2 to 2.6. The rate differs from country to country which depends also on the preventative strategy applied by the government of that country. Comparing to COVID-19, the reproduction rate for seasonal flu is just about 1 that means an infected person will continue to infect another 1 or 2 persons only. SARS has reproduction rate at 3 which is more contagious and more deadly than COVID-19. SARS is very contagious only from the second week onwards after the symptoms appeared. In contrast, COVID-19 patient who first contracted the decease will become contagious immediately, symptoms then will appear 1 or 2 days later. Between the time prior to 48 hours, he got the symptoms and went to the hospital to be tested, he could potentially infect anybody that he will be in contact with. By using the S.E.I.R. model, it shows that this latency period is dangerous and R_0 can be subsided significantly by social distancing. The following screen-captures show the progression of COVID-19 spread, in the cases of 0 isolation, 80% isolation and 100% isolation, in Fig. 3.16. Mainly it shows how important isolation is. The simulation is programmed and hosted on a website, with URL: https://www.trackcorona.live/isolation. The simulation shows how self-isolation, enforced quarantine for suspected patients and social distancing contribute to 'flattening the curve' of COVID-19 pandemic.

In the near future, it is anticipated that the modelling of virus pandemic behaviours would be aided by AI for more precise estimating the ideal lockdown period. It is known that too short the period, the pandemic may re-bounce, too long the lockdown will harm deeper the economy driving high the unemployment rate and recession,

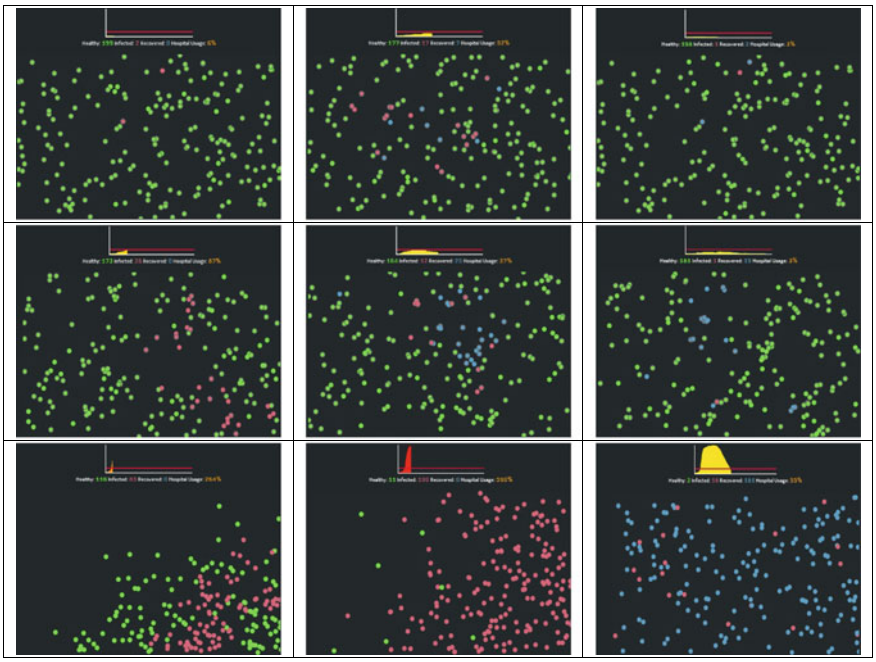

Fig. 3.16 Simulation of COVID-19 pandemic progress by S.E.I.R. model with (top) 100% social distancing, (middle) 80% social distancing, (below) no social distancing

etc. It could be formulated as a multi-objective optimization model by AI, in such a way that the death toll, the infected rate, the economy damage, the lockdown period and the necessary resources to fight the virus be minimized, but the number of susceptibles be maximized, saving life at the lowest costs, choosing intelligently the right mix of preventative strategies.

AI can also play a significant role in predicting more accurately the outcomes than simple mathematics models, being able to simulate the behaviours of the pandemic with more details using supercomputers. A recent work on decision support by stochastic simulation for resource allocation to fight COVID-19 is published [18]. Large-scale simulation by AI and superior computing hardware could be conducted along this direction. Moreover, AI has been used for drug discovery, drug design and drug testing which helps speeding up the R&D of COVID-19 vaccine [19].

References

1. Bowles J (2020) How Canadian AI start-up BlueDot spotted Coronavirus before anyone else had a clue, Diginomica, March 10, 2020, Accessed 18 April 2020
2. Nagypál G (2005) Improving information retrieval effectiveness by using domain knowledge stored in ontologies. In: Meersman R, Tari Z, Herrero P (eds) On the Move to Meaningful

Internet Systems 2005: OTM 2005 Workshops. OTM 2005. Lecture Notes in Computer Science, vol 3762. Springer, Berlin, Heidelber
3. Mohd Sharef N, Kasmiran KA (2012) Examining text categorization methods for incidents analysis. In: Chau M, Wang GA, Yue WT, Chen H (eds) Intelligence and security informatics. PAISI 2012. Lecture Notes in Computer Science, vol 7299. Springer, Berlin, Heidelberg
4. Tosey P, Mathison J (2009) What is NLP? The 'Six Faces' of the field. In: Neuro-linguistic programming. Palgrave Macmillan, London
5. Mohammed M, Hissam A, Mohammed T, Anya SO, Applications of big data analytics: trends, issues, and challenges, Springer, 2018. ISBN: 978-3-319-76471-9
6. Tromblay DE, Spying: Assessing US Domestic Intelligence Since 9/11, Lynne Rienner Publishers, February 25, 2019, ISBN: 978-1626377806
7. Alessa A, Faezipour M (2018) A review of influenza detection and prediction through social networking sites. Theor Biol Med Model 15:2. https://doi.org/10.1186/s12976-017-0074-5
8. Jiang X, Coffee M, Bari A, Wang J, Jiang X et al (2020) Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. CMC-Comput Mater Continua 63(1):537–551
9. Fong S, Deb S, Yang X, Li J (2014) Feature selection in life science classification: Metaheuristic Swarm search. IT Professional 16(4), 24–29
10. Bhadra P, Yan J, Li J et al (2018) AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. Sci Rep 8:1697. https://doi.org/10.1038/s41598-018-19752-w
11. Greene T (2020) Alibaba's new AI system can detect coronavirus in seconds with 96% accuracy, TNW, 2 March 2020. Accessed 19 April 2020
12. Boyd A (2019) Military Algorithm Can Predict Illness 48 Hours Before Symptoms Show, Next Government, 24 Oct 2019. Accessed 19 April 2020
13. Fong SJ, Li G, Dey N (2020) Rubén González Crespo, Enrique Herrera-Viedma, Finding an accurate early forecasting model from small dataset: a case of 2019-nCoV novel coronavirus outbreak. IJIMAI 6(1):132–140
14. Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. proceedings of the royal society a: mathematica., Phys Eng Sci 115 (772): 700. Bibcode:1927RSPSA.115..700 K. https://doi.org/10.1098/rspa.1927.0118.JSTOR94815
15. Vynnycky, Emilia; White, Richard G. An Introduction to Infectious Disease Modelling. Retrieved 2016–02-15. An introductory book on infectious disease modelling and its applications
16. Aron JL, Schwartz IB (1984) Seasonality and period-doubling bifurcations in an epidemic model. J Theor Biol 110:665–679
17. Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner N, Ke R (2020) High contagiousness and rapid spread of severe acute respiratory syndrome Coronavirus 2, EID Journal 26(7), 2020, ISSN: 1080-6059
18. Fong SJ, Li G, Dey N, Crespo RG, Herrera-Viedma E (2020): Composite monte carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. Appl Soft Comput 9:106282
19. Le TT, Andreadakis Z, Kumar A, Román RG, Tollefsen S, Saville M, Mayhew S (2020, The COVID-19 vaccine development landscape, Nature Reviews Drug Discovery, 9 April 2020, ISSN 1474-1784