Literature Review

Applying Machine Learning to improve the prediction of future COVID-19 cases in Ireland

Meet Mehta

Patrick McLoughlin

Simon McLain

February 13, 2022

Table of Contents

**Contents**

[**1.** **Introduction** 4](#_Toc95072533)

[**1.1.** **Mathematical & Statistical Models** 5](#_Toc95072534)

[**1.1.1.** **Likelihood Estimation to predict Reproduction Number** 5](#_Toc95072535)

[**1.1.2.** **SEIRD model** 6](#_Toc95072536)

[**1.1.3.** **Compartmental model** 6](#_Toc95072537)

[**1.2.** **Big Data** 8](#_Toc95072538)

[**1.2.1.** **Social Media Big Data** 9](#_Toc95072539)

[**1.2.2.** **Healthcare Big Data analytics** 10](#_Toc95072540)

[**1.2.3.** **Demographic Big Data** 11](#_Toc95072541)

[**1.3.** **Machine Learning** 11](#_Toc95072542)

[**6.** **Conclusion** 12](#_Toc95072543)

[**Works Cited** 12](#_Toc95072544)

Table of Figures

# **Introduction**

COVID-19, also known as the novel coronavirus disease which is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV2), is a highly contagious sickness that the World Health Organization (WHO) has labelled a global pandemic. COVID19 is a zoonotic coronavirus that is related to the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) and SARS-CoV-2 viruses that were seen in previous decades. The virus has a high infectivity rate and a high morbidity rate in elderly persons and those with serious illnesses like asthma, cancer, and diabetes. More than 360 million confirmed cases and 5.6 million deaths had been reported worldwide as of Jan 26, 2022. In Ireland, from 3 January 2020 to 5:30pm CET, 26 January 2022, there have been 1,153,666 confirmed cases of COVID-19 with 6,087 deaths, reported to WHO. As of 8 January 2022, a total of 9,849,271 vaccine doses have been administered. Every nation in the world is affected by the COVID-19 virus and has suffered significant social and economic loss at a global scale. The deteriorating situation necessitates the use of containment techniques as soon as possible to prevent the spread of the disease. Because there is currently no therapy or medicine for the virus, effective planning of health services and infrastructures is critical.

Administrators and public health professionals are under a lot of pressure to manage the accommodation of COVID19 patients. As a result, certain forecasting methods will be necessary to estimate the number of new COVID-19 cases that may emerge in the near future, as well as to organise the resources and equipment needed to combat the outbreak. Public health officials may be able to use illness prediction in advance to plan for the effective and timely allocation of resources needed for medical treatment in the event of a pandemic.

To tackle this issue, several mathematicians and data scientists are working to create the model that predicts accurate COVID-19 cases in different countries. This literature review sheds light on numerous predictions models adapted by different researchers all over the world to select the optimal model for predicting covid cases specifically in Ireland. The model will predict data with minimum error and thus will prove critical in making decisions for the government. The outcomes from this study will drastically improve the quality of health care system and management. Health care system can perform much better knowing the future covid cases and cater to the needs of public. This research compares such models presented by research papers and include the information on several COVID-19 prediction models, the characteristics used while building models, the accuracy, error rate and the result obtained when model was applied in real life scenario.

Following inclusion and Exclusion criteria were used

Inclusion Criteria

• Papers related to prediction of COVID-19 using Machine Learning algorithms.

• All articles should be in English language.

Exclusion Criteria

• Incomplete articles.

# **Mathematical & Statistical Models**

Mathematical models are widely popular in predicting because of their ability to track and predict the course of an epidemic in a variety of situations. Mathematical models are based off equations that determine how the system impacts from a change in one variable and its resulting impact on other variables.

# **1.1.1.** **Likelihood Estimation to predict Reproduction Number**

**Shi Zao** [1]published the paper that estimates the unreported number of Covid-19 cases in first half of January. The paper present a novel methodology for predicting the time evolution of the cumulative number, N(t), of the individuals reported to be infected in a given country. The research paper suggests that there were number of cases unreported in January.

The unreported cases were denoted by *ξ* under-reported in January. Cumulative total cases were denoted by *Ci* which would be equal to the sum of cumulative cases reported *ci*, and cumulative unreported cases *Ξi.* Thus the mathematical equation was

*Ci* = *ci* + *Ξi*.

*Ci*  : Total cases

*ci*  : Reported cases

*Ξi.* : Unreported cases

The number of unreported cases and the basic reproduction number, R0, of 2019-nCoV was assessed from 1st Jan to 15th Jan 2020, along with the number of unreported cases. Method of exponential growth was used to estimate the epidemic curve of 2019-nCoV cases in mainland China from 1st Dec 2019 to 24th Jan 2020. Maximum likelihood estimation was used to determine the number of unreported instances. To estimate R0, serial intervals (SI) of infection induced by two other well-known coronaviruses (CoVs), Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS) were used. The initial expansion phase was verified to have followed an exponential development trend. This result proved that under-reporting was almost certainly the cause and massive amount if cases were not reported. Between 1st Jan and 15th Jan, 2020, there were likely 469 (95 percent CI: 403–540) unreported cases due to under-reporting. In comparison to the average condition from 1st Jan to 17th Jan, 2020, the reporting rate was predicted to have increased 21-fold (95 percent CI: 18–25). The R0 of 2019-nCoV was calculated to be 2.56 (95 percent confidence interval: 2.49–2.63). The prediction of these numbers proved that the under-reporting most likely occurred in the first half of January 2020 but there was no way to confirm these data with actual numbers.

# **1.1.2.** **SEIRD model**

A screenshot of a computer

Description automatically generated with low confidence

Fig 1. Schematic of modified SEIRD model

SEIRD is a five-compartment epidemic model which stands for Susceptible (S)-Exposed (E)-Infected (I)-Recovered (R)-Death (D). This model has been frequently used to predict the dynamics of contagious diseases.

**Vipin, Namrata and Nandan** [2]utilised the SEIRD model to predict COVID-19 cases in India under the effect of lockdown. They created improvised SEIRD model with the modified transmission rate [β/k] based on intervention strategies like lockdown, time dependent reproduction number [R (t)] and under reported cases. The data up to 10th July, 2022 was analysed and suggested that lockdown was controlled by 50% by implementing lockdown in India at the beginning of transmission of disease phase. Results proved that effective reproduction number has been varied within range of 2-3 during lockdown period. The research estimated the COVID-19 epidemic peak under different possible lockdown scenarios. The research predicted the maximum number of total confirmed cases and total active cases was around 2.5 × 106 and 5.8 × 105, respectively on 15th August, 2020. The study further suggests that lockdown should have been extended to 100 days (after 65 days of initial lockdown) to better control the transmission of COVID-19.

# **1.1.3.** **Compartmental model**

**Tang et al.**[3]proposed a mathematical model to determine the transmission rate of COVID-19 to predict the confirmed cases in the next seven days. The research states that basic transmission number can be estimated using mathematical modelling to help determine the potential and severity of an outbreak, also it can provide crucial information for deciding the type and intensity of disease management. Based on the clinical development of the virus, the epidemiological condition of the individuals, and intervention methods, a deterministic compartmental model was developed. The control reproduction number might be as high as 6.47 (95 percent CI 5.71–7.23), according to estimates based on likelihood and model analysis. Sensitivity analyses show that interventions like intensive contact tracing preceded by quarantine and isolation can effectively reduce the control reproduction number and transmission risk. The model made a very good prediction of the confirmed cases from 23rd to 29th January 2020. **1.1.4. ARIMA model**Diagram

Description automatically generated

Fig 2. Schematic of ARIMA model

ARIMA is one of the most used time series models as it takes into account changing trends, periodic changes and random disturbances in the time series. The ARIMA model is a class of statistical models for analysing and forecasting time series data. ARIMA (p, d, q) where p, d, q are the order of autoregressive, the degree of differencing, and the order of moving average respectively.

**Ceylan** [4] applied ARIMA to predict the prevalence of COVID-19. Time series models are critical for anticipating the impact of the COVID-19 epidemic and implementing the necessary countermeasures. Auto-Regressive Integrated Moving Average (ARIMA) models were built in this research to forecast the epidemiological trend of COVID-19 prevalence in Italy, Spain, and France. The key aspects of the model include AR : Autoregression, I : Integrated and MA : Moving Average. Each of these component are explicitly specified in the model as a parameter, which gives the standard notation of ARIMA(p,d,q). Different ARIMA parameters were used to create several ARIMA models. For Italy, Spain, and France, the ARIMA (0,2,1), ARIMA (1,2,0), and ARIMA (0,2,1) models with the lowest MAPE values (4.7520, 5.8486, and 5.6335) were chosen as the best models. This study demonstrates that ARIMA models are capable of forecasting COVID-19 prevalence in the future.

# **Big Data**

Big Data is defined as “data that contains greater variety, arriving in increasing volumes and with more velocity” [5]. In the past data was stored mostly on paper records of local digital storage devices, if at all. As such the availability of data was much reduced when compared with today. In essence the dramatic increase in data collection then storage in a shareable way as potentially provides scientists a wealth of insights. With this technological leap scientists can enrich the data they collect themselves through observations and experiments with open-source data available collected from a multitude of organisations and institutions.

Epidemiologists undertake systematic, data-informed, investigations into the biology of pathogens and their complex interactions with their hosts to identify changes in behaviour that could contain an outbreak reduce or even eliminate the spread of disease. Big Data can potentially be utilised to enrich their knowledge and help build reliable predicts of the development of an epidemic as well as the effectiveness of control measures.

Potentially valuable Big Data artifacts can be broadly placed in three categories.

Firstly, healthcare Big Data relevant to our project will include detailed information collected from patients who have presented with disease. It will provide information related to age, gender and may be enriched with data relating to medical health, pre-existing conditions, vaccination status etc. This information is of course collected after outbreaks of the disease have occurred, but it is specific to the disease as it will contain information for individuals who have been proven to be infected. The data is normally collected by health care providers and government agencies. This data is useful for identifying risk factors such as age and gender. As it is collected in real time during an outbreak situation the data will evolve over time, as more knowledge of the disease is gained.

Other healthcare datasets can also be of use, for instance, if certain underlying conditions (e.g., cardiovascular disease) are risk factors for diseases, such as COVID-19, becoming life threatening then knowledge of their prevalence in the community may also clarify the severity of the disease.

The next relevant Big Data category is s population demographics. Reliable demographic information will typically be collected by governments through a census. It describes the population geospatially, by ethnicity, religion, gender, and age. Of course, this data is not specific to the outbreak of a disease, however it can provide valuable contextual information. For instance, demographic data can help to understand population density, a factor that will be important for the spread of airborne diseases. In normal circumstances this data will not change significantly over a period of several months. So, the most recent census data can be taken as generally representative of the composition of the population.

The final relevant Big Data category is Social Media data. This data can come from various sources such as, Google Searches, posts to Twitter, Facebook, Instagram etc. This data is typically unstructured, taking the form of text, emoji, gifs, pictures, and documents. Social media data has the advantage of being available in real time and being representative of the cognizance and behaviours of the actors participating within the Social Media space. Relevant applications in the context of the spread of COVID-19 could be Google searches for testing centres, pharmacists, and symptoms of the disease. The behavioural indicators can include hashtags that indicate compliance or non-compliance with public health advice, as well as number of people contained within photographs. As such this data may be helpful in understanding whether or not a community is engaging in behaviours that would increase or inhibit the spread of disease.

# **Social Media Big Data**

[6] Use data from Google Trends and Twitter to detect influenza in Greece. Hardware included a computer with uninterrupted power supply. An ARIMA(X) and a custom prediction model were coded in the Python language. Their research demonstrated that internet data can be useful in tracking epidemics with the potential to estimate and predict influenza. Twitter REST API and the Twitter Streaming API can be used to gather tweets, but are rate limited reducing the amount of data that can be gathered. Google data can be accessed through their web application as well as using API’s such as PyTrends.

[7] propose improve predictions by using neural network based Long Short-Term Memory models and auto-regressive moving average models with two channels of inputs to incorporate social media and historic disease data. The keras2 python library is used to implement the machine learning algorithm. The approach concluded that social media data can be used in combination with current case counts to improve the accuracy of prediction and proposed enriching the dataset with addition information such as climate data.

[8] Gathered and pre-processed English language and Tagalog language Flu related tweets through the Twitter API. Pre-processing included removal of punctuations, special characters, hashtags. Programming was carried out using the Python programming language and associated libraries. Including the use of the Natural Language ToolKit [NLTK] for stemming and lemmatisation. Machine Learning algorithms were implemented using the Scikit Learn library. Their goal was to apply the C5.0 decision tree algorithm to improve on predictions made using the Naïve Baynes algorithm. Their implementation of the C5.0 model had a significant increase in accuracy when compared to Naïve Baynes, but was more complex in terms of Big O.

[9] Discuss the failure of researchers at Google to use Big Data to accurately predict flu outbreaks, particularly how the model was initially accurate but then began to perform worse over time. An example citing being the algorithms’ weakness to overfitting on seasonal factors unrelated to influenza, such as “high school basketball”. In their research the authors highlight both the benefits and challenges of using Big Data for the public good. They do this in terms of the selection and pre-processing of the data so that it contains components with a direct, real world, association with the dependent variable as well as the general challenges accessibility to data and the need for companies and institutions to share data for the public good whilst protecting the data subjects right to privacy.

[10] Researches the use of a popular Chinese micro-blog called ‘Sina’ to predict the health status of citizens in a region of Beijing. Users of the micro-blog can make posts of 140 characters or less. Data was accessible via an API and the researchers were able to limit the data source to nearby users. They categorised 5000 posts into sick and not-sick for their training and test set. Thus, adopting a supervised learning approach. Programming was carried out in Python. Pre-processing included word-weighting utilising Boolean weighting, term frequency weighting, inverted document frequency weighting and term frequency-inverted document weighting frequency. A classification approach utilising K-Nearest Neighbour and Support Vector Machine algorithms were tested. Their algorithm was able to predict an outbreak of influenza 5-days before the Chinese national influenza centre.

[11] Demonstrates the used of multiple Big Data sources to perform a retrospective study of prediction of outbreaks of COVID-19. Taking a two-step approach, first step being a lag correlation analysis to find the number of days ahead to make an accurate prediction. Secondly to apply Machine Learning algorithms. The methodology applied was as follows: four Big Data sources, social media, web-search, air quality and daily COVID-19 new cases were pre-processed, underwent Log Correlation Analysis, and then fed into Machine Learning and Deep Learning based prediction models. The pre-processed web-based social media and search data being sorted into data collected during lockdown/ unlock phases and that collected outside of nationwide lockdowns. A regression approach was taken for the Machine Learning approach, utilising the Support Vector Regression and polynomial algorithms. Several Deep Learning algorithms were used: multi-layer perceptron, ElasticNet and deep neural network. They conclude that daily new cases could be accurately predicted 70 – 100 days ahead with the inclusion of air quality data improving the performance of the models.

[12] The method proposed by these researchers is again a hybrid-model including data source from the United States’ Centre of Disease Control [CDC] and flu-related data sourced from Twitter. The objective of the paper is to predict future influenza activities, to do so more accurately than current methods, and to combine the real-time feed from Twitter with historical data from the CDC. Using the most recent five weeks of data from both sources to produce a model that accurately predict influenza 2-3 weeks ahead. Their model improved upon the predictions of the 2013 version of Google Flu Trends.

[13] This study, focusing on sub-Saharan Africa, identifies two important trends. Firstly, that searches on the world-wide-web for health-related information can be helpful in monitoring disease trends, and the terms of the searches are specific to sub-Saharan Africa, insofar as they may not be predictive of influenza trends in other countries. Secondly, they highlighted the need for greater public health education regarding diseases to aid the citizens in conducting safe and productive internet searches. The researchers conducted a hybrid approach using data from Cameroon’s surveillance system ILI as well as Google Search data. They evaluate several multivariate regression models, the Random Forrest ensemble model and ARIMA.

[14] Propose Hadoop as the pre-processing platform for Social Media Big Data Analytics. Data is sourced from Twitter, and whereas Relational Database Management Systems are traditionally used the authors propose using Hadoop’s Map Reduce Model to transform the data into key-value pairs to increase the data volume, processing capability, deal with structured and unstructured data, for producing ad hoc reports, fault tolerance, all in a more cost-effective platform. Apache Hive is used for querying the data and real-time tweets are streamed from Twitter using an API. Their methodology aimed to analyse Tweets within the minimum time.

# **Healthcare Big Data analytics**

[15] Investigate Big Data Analytics in healthcare, studying its unique characteristics, different analytics phases, platforms, and tools used, challenges and their conclusions. Ending with a discussion of its further evolution. The unique characteristics, commonly referred to as V’s, describe the data in four dimensions: Volume, Velocity, Variety and Veracity. Volume relates to the amount of data. Velocity relates to the rate at which the data is created. Variety relates to the state of the data: structured, unstructured, and semi-structured. Finally, Veracity relates to the trustworthiness that the data is suitable and valuable for the purpose in which it will be used. They conduct a literature review of the existing systems in which data is stored and used. Their analysis summaries to exiting system and proposes improvements. The weakness of the existing systems being described as its inability to support live streaming, limiting the availability of the data. Their proposal lays out a technology for improving data availability for its present architecture using the MapReduce as well as proposing the use of The Spark and Kafka technology for a new big data framework.

[16] Discuss the processes that make Big Data in health care different from nominal data. They break it down in terms of features [Structured Features, Linked Data, Multisource and Multiview Data, Streaming Data and Features, Scalability and Stability], challenges in Big Data representation and reduction, then discuss the merits of a variety of potential classifications of Machine Learning algorithms. They discuss the variety of data stored in electronic health records includes health data, demographics, medical history and more, that can be collected using techniques such as interviews, sensors, laboratory analysis and IOT devices. Identifying that before Machine Learning techniques are applied practical issues need to be overcome in how this data is networked, secured, and distributed. The data pre-processing steps include data organisation and structuring, data cleaning and reduction, data integration and processing. Thereafter, selecting, tuning, and evaluating a variety of algorithms can occur.

# **Demographic Big Data**

[17] Investigate the use patient level data using the New York State Department of Health Statewide Planning and Research Cooperation System. Their method combines this data with high resolution spatial and demographic analysis techniques to assess census-tract level variation in breast cancer screening and diagnosis rates. Geo-spatial analysis was used using the addresses of target patients and population data from the American Community Survey of the US census survey. They found significant differences in the rate of Breast Cancer screening by race and discussed variances is diagnosis and mortality rates across different race and ethnic groups. Concluding that incorporating demographics into predictive models can assist in a targeted approach to help address these disparities. This is of interest as a comprehensive model of predicting COVID-19 cases could potentially account for demographic factors that influence the transmission of the virus,

[18] This study seeks to implement baseline calculations for the prediction of demographic markers to identify demographic markers in photos, biographies and tweets. Taking a bag-of-words approach to feature selection and then evaluating several machine learning models. In their model they find it easier to predict age and gender than age and location. Even so, they claim their approach has advantages, namely addressing bias present in hard-coded dataset and eliminating the effort required to hard-code datasets for a Corpus. Such a method could be helpful for aligning information contained within Social Media posts with demographic data gathered from Census data, or example to compare the extent to which the social media posts represent the overall population in terms of age, gender, ethnicity, race, etc.

# **Machine Learning**

Machine learning (ML) is the study of having computers replicate the behaviours of human learning to identify and attain knowledge, skills and continually improve upon performance, as described by [19] with [20] reinforcing this, but further stating that ML uses new knowledge to advance its performance. [21] says, “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”. An example of this, if we want an algorithm to classify symptoms of patient as having a disease or not. Experience E would be the symptoms already classified as a disease with Task T automatically assessing and classifying new patients and their symptoms as having a disease. Performance P should increase the classifiers rate of accuracy when determining the disease based on the symptoms.

Due to having reason, when a human has an experience, they learn from it whereas, computers learn from algorithms. With the availability of datasets that are large and complex that prevents normal processing from humans, and with computational power growing, ML is used to take a sophisticated approach using various algorithms to find value in data.

There are many ML algorithms, but they are broken into 4 main classifications based upon the approaches they take. These classifications are supervised, unsupervised, semi supervised and reinforcement learning.

Supervised learning uses labelled data for training to make predictions for the response variable which can be categorical or continuous. Without prior labelling, unsupervised learning learns from the data and makes classifications without any know priors. When not all data is known semi-supervised learning is used where there is some labelled data with the majority being unlabelled. Finally, when it’s known what is wanted but it’s not known how to get it then reinforcement learning is used which maximises the reward in certain situations to find the best behaviour.

# **Conclusion**

<< This is a critical part of the literature review. Briefly state again the overall context of the work. Briefly state the literature review process including technical core areas and subcategories. Compare and contrast the relevance of the technologies under discussion. Compare and contrast the potential for existing technologies to address the overall research challenge. Identify limitations in existing approaches. Discuss future work which can enhance the existing state of the art in order to address the overall research challenge>>

# **Works Cited**

|  |  |
| --- | --- |
| [1] | S. Zhao, Estimating the unreported number of novel coronavirus (2019-ncov) cases in china in the first half of january 2020: a data-driven modelling analysis of the early outbreak, Wuhan: Journal of clinical medicine, 2020. |
| [2] | Vipin Tiwari, Mathematical Modeling Based Study and Prediction of COVID-19 Epidemic Dissemination Under the Impact of Lockdown in India, Frontiers in Physics, Vol : 8, 2020. |
| [3] | B. Tang. e. al., Estimation of the Transmission Risk of the 2019-nCoV and Its Implication for Public Health Interventions, vol. 9, Journal of Clinical Medicine, Feb. 2020, p. 462. |
| [4] | Z. Ceylan, Estimation of covid-19 prevalence in italy spain and france, Science of The Total Environment, 2020. |
| [5] | Oracle, “What is Big Data,” [Online]. Available: https://www.oracle.com/big-data/what-is-big-data/. [Accessed 31 01 2022]. |
| [6] | L. Samaras, E. Garcia-Barriocanal and M.-A. Sicilia, “Comparing Socila media and Google to detect and predict severe epidemics,” Springer Nature, 2020. |
| [7] | A. Parwez, M. Abulaish and Jahiruddin, “A Social Media Time-Series Data Analytics Approach for Digital Epidemiology,” in *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 2020. |
| [8] | L. Albancdes, B. A. Bungar, J. P. Patio, R. J. M. Sevilla and D. Acula, “Application of C5.0 Algorithm to Flu Preduction Using Twitter Data,” in *2018 International Conference on Platform Technology and Service (Platcon)*, Manila, 2018. |
| [9] | D. Lazer and R. Kennedy, “What Can We Learn from the Epic Failure of Google Flu Trends,” Wired, 1 October 2015. [Online]. Available: https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/. [Accessed 02 2 2022]. |
| [10] | N. Yang, X. Cui, C. Hu, W. Zhu and C. Yang, “Chinese Social Media Analysis for Disease Surveillance,” in *2014 International Conference on Identification, Information and Knowledge in the Internet of Things*, Wuhan, 2014. |
| [11] | S. Chatterjee, K. Ghosh, A. Banerjee and S. Banerjee, “Forecasting COVID-19 Outbreak Through Fusion of Internet Search, Social Media, and Air Quality Data: A Retrospective Study in Indian Context,” IEEE, 2020. |
| [12] | K. Lee, A. Agrawal and A. Choudhary, “Forecasting Influenza Levels using Real-Time Social Media Streams,” in *IEEE International Conference on Healthcare Informatics*, Evanston, 2017. |
| [13] | E. Nsoesie, O. Oladeji, A. Abah Abah and M. L. Ndeffo-Mhab, “Forecasting influenza-like illness trends in Cameroon using Google Search Data,” Nature, 2021. |
| [14] | F. Shaikh, F. Rangrez, A. Khan and U. Shaikh, “Social Media Analytics Based on Big Data,” in *International Conference on Intelligent Computing and Control*, Mumbai, 2017. |
| [15] | G. Sasubilli and A. Kumar, “Machine learning and Big Data Implementation on Health Care data,” in *International Conference on Intelligent Computing and Control Systems*, 2020. |
| [16] | P. Saranya and P. Asha, “Survey on Big Data Analytics in Health Care,” in *Smart Systems and Inventive Technology*, 2019. |
| [17] | F. Murphy, K. Abel-Hart and F. Wang, “A Fine-grain Geospatial and Demographic Analysis of Breat Cancern Patterns in New York State,” in *IEEE International Conference on Big Data*, 2021. |
| [18] | J. Radford, L. Horgan and D. Lazer, “Baselines for Demographic Inference on a New Gold Standard Twitter Corpus,” in *IEEE International Conference on Big Data*, Boston, 2017. |
| [19] | T. Hengyao and L. Haoyong, “Machine Learning Methods And Their Application Research,” in *2011 International Symposium on Intelligence Information Processing and Trusted Computing* , Wuhan, 2011. |
| [20] | I. Portugal, P. Alencar and D. Cowan, “The use of machine learning algorithms in recommender systems: A systematic review,” *Expert Systems with Applications,* pp. 205-227, 2016. |
| [21] | R. Michalski, J. Carbonell and T. Mitchell, “Machine learning: An artificial intelligence approach,” *Artificial Intelligence,* vol. 25, no. 2, pp. 236-238, 1985. |