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](#_Toc95588615)

[**1.1.** **Mathematical & Statistical Models** 5](#_Toc95588616)

[**1.1.1.** **Likelihood Estimation to predict Reproduction Number R0** 5](#_Toc95588617)

[**1.1.2.** **SEIRD model** 6](#_Toc95588618)

[**1.1.3.** **Compartmental model** 6](#_Toc95588619)

[**1.1.4.** **ARIMA model** 7](#_Toc95588620)

[**1.2.** **Epidemiological Modelling of COVID-19 in the Republic of Ireland** 8](#_Toc95588621)

[**1.2.1.** **Time-Series Estimation of Effective Reproduction number** 8](#_Toc95588622)

[**1.2.2.** **Epidemiological parameters** 8](#_Toc95588623)

[**1.2.3.** **Monitoring Country Level Incidence** 9](#_Toc95588624)

[**1.2.4.** **Population-level SEIR** 9](#_Toc95588625)

[**1.3.** **Big Data for predictive modelling** 9](#_Toc95588626)

[**1.3.1.** **Predictive Modelling with Social Media Big Data** 10](#_Toc95588627)

[**1.3.2.** **Predictive Modelling with Healthcare Big Data** 11](#_Toc95588628)

[**1.3.3.** **Demographic Big Data and predictive modelling** 12](#_Toc95588629)

[**1.4.** **Machine Learning** 12](#_Toc95588630)

[**1.4.1.** **Supervised Learning** 13](#_Toc95588631)

[1.4.1.2 Regression 13](#_Toc95588632)

[1.4.1.3 Classification 14](#_Toc95588633)

[**1.4.2.** **Unsupervised Learning** 15](#_Toc95588634)

[**2.1** **Conclusion** 15](#_Toc95588635)

Table of Figures

[Figure 1. Schematic of modified SEIRD model 6](#_Toc95588844)

[Figure 2. Schematic of ARIMA model 7](#_Toc95588845)

[Figure 3. IEMAG Technical Notes 8](#_Toc95588846)

[Figure 4 - Polynomial Regression Formula 13](https://studentait-my.sharepoint.com/personal/a00279668_student_ait_ie/Documents/Literature%20Review/Literature%20Review.docx#_Toc95588847)

[Figure 5 - Linear Regression Formula 14](#_Toc95588848)

# **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 and the data sources utilised by different groups of 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.
* Papers related to use of Social Media Data to predict cases of COVID-19 and influenza.
* Papers related to the use of Healthcare Big Data.
* Technical Notes published by the Irish Epidemiological Advisory Group
* Papers related to the incorporation of population demographics in Machine Learning algorithms.
* All articles must 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. Statistical models on the other hand are the mathematical model that embodies a set of statistical assumptions to generate sample data and make predictions about the real world.

# **1.1.1.** **Likelihood Estimation to predict Reproduction Number R0**

**Shi Zao** [1]published the paper that estimates the unreported number of Covid-19 cases in first half of January. The paper presents 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 i.e. Reproduction number, 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

Figure 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

Figure 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.

# **Epidemiological Modelling of COVID-19 in the Republic of Ireland**

The Chief Medical Officer and the National Public Health Emergency Team (NPHET) for the Republic of Ireland is provided with statistical and mathematical support by the Irish Epidemiological Modelling Advisory Group (IEMAG) [5] who publish technical notes, intended to be read by practitioners from a range of scientific disciplines, which in a series of technical notes through 2020 – 2021 outline their approach to estimating and modelling of subjects relevant to the spread of COVID-19 in Ireland.

|  |  |
| --- | --- |
| Title | Date Published |
| A population-level SEIR model for COVID-19 scenarios | 11/05/2020 |
| A Hierarchical Model for Monitoring Country Level Incidence: IEMAG Epidemiology Modelling Subgroup | 27/05/2020 |
| Time-Series Base Estimation of the Effective Reproduction Number (Rt) | 21/05/2020 |
| COVID-19 epidemiological parameters summary document | 20/05/2020 |
| COVID-19 Hospital Utilisation Planning model: IEMAG Service Demand Modelling Subgroup | 24/09/2020 |
| A population-level SEIR model for COVID-19 scenarios (updated technical note) | 12/11/2021 |
| COVID-19 Hospitalisation Utilisation Planning model (updated technical note) | 25/02/2021 |
| Hospital admission probability and length of stay among Covid-19 confirmed cases | 18/02/2021 |
| COVID-19 Hospital Utilisation Planning Model (updated technical note) | 23/04/2021 |
| Hospital Admission Probability and Length of Stay among COVID-19 confirmed cases | 30/04/2021 |

Figure 3. IEMAG Technical Notes

The technical notes describe the statistical approach taken by IEMAG to model the epidemiology of COVID-19 and that upon which their estimations where used.

# **Time-Series Estimation of Effective Reproduction number**

[6] The IEMAG utilise the same approach as [7] to estimate Rt which provides an estimation method when only the number of daily cases is known and there is little of no contact tracing information available that would enable an estimation of the effective reproduction number. If Rt is less than 1.0 then the virus is not spreading and will ultimately disappear. If Rt is greater than 1.0 means that the virus is spreading fast [8]. Rt difference from R0 in so far as R0 is the initial reproduction of the virus as the beginning of a pandemic and Rt refers to a particular point in time. This approach does not predict how many cases will occur in the future, rather it helps to predict the actual number of cases occurring in the population based on the number of confirmed cases. Rt can be helpful in determining the effectiveness of lockdown measures.

The following Healthcare data was used to calculate Rt:

* HPSC 24-hour external tests backdated.
* Confirmed cases, incidence by specimen collection date.
* COVID-19 Hospital admissions.

# **Epidemiological parameters**

The following parameters were included in the COVID-19 transmission model for Ireland, being selected from documents relevant to the virus’s parameterisation. [9]

* R0/R
* Latent period & relative importance of pre-symptomatic period
* Incubation period
* Generation time & serial interval
* Proportion of infected who are asymptomatic, by age
* Length of infectious period in asymptomatic people an in symptomatic people who do not isolate
* Time from onsite of symptoms to diagnosis/ test results and to hospitalisation
* Length of hospital stay and admission to ICUs
* Relative infectiousness of asymptomatic versus symptomatic infected people

# **Monitoring Country Level Incidence**

[10] describes the application of a Bayesian hierarchical Poisson model fitted to the daily case data from the Health Protection Surveillance Centre’s [HPSC] Computerised Infectious Disease Reporting [CIDR] Database proposed to be used for monitoring country level daily cases per 100,000 population. Peak incidence timing and peak incidence size is also estimated, in terms of daily new cases. Their model is fitted using the Stan programming language, which interfaces with popular data analysis programming languages such as R, Python, MATLAB etc. [11].

# **Population-level SEIR**

[12] conducted in 2020 and based on a contemporaneous literature review of the then understanding of the epidemiology of COVID-19 the authors used the CIDR database to fit a Bayesian Poisson model with rate following a Gompertz function. Programming is implemented in the Stan programming language interfacing with the R programming language. R is a popular programming language for statistical analysis [13]. Their work describes several modelling scenarios where their model infers R and daily new cases as a function of days from a specified day.

# **Big Data for predictive modelling**

Big Data is defined as “data that contains greater variety, arriving in increasing volumes and with more velocity” [14]. 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 potentially provides scientists and non-scientists alike a wealth of information to explore.

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.

Through the literature view Big Data potentially valuable for epidemiological investigatory purposes could 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, often 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, and/ or are searching for resources related to the treatment and care of disease.

# **Predictive Modelling with Social Media Big Data**

[15] Use data from Google Trends and Twitter to detect influenza in Greece. Hardware included a computer with uninterrupted power supply. The ARIMA(X) Machine Learning model 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.

[16] propose improving 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.

[17] 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.

[18] 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’ rights to privacy.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

# **Predictive Modelling with Healthcare Big Data**

[24] 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.

[25] 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 and predictive modelling**

[26] 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,

[27] 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 [28] with [29] reinforcing this, but further stating that ML uses new knowledge to advance its performance. [30] 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.

# **Supervised Learning**

## Regression

Regression is a way of determining which independent variable(predictor), denoted x, does have an impact on the dependent variable(target), denoted y, and by finding this relationship, allows one to predict the effects x has on y.

[31] carried out a comparative analysis on future COVID cases, deaths, recoveries for the next 5 days in India using 6 regression models which were Polynomial Regression(PR), Decision Tree(DT), Linear Regression(LR) and Random Forest(RF), Support Vector Machine(SVM), Least Absolute Shrinkage and Selection Operation(LASSO). The parameters that were used in this study to measure the results were R2 score, Mean Absolute Error(MAE), Root Mean Square Error(RMSE), Mean Square Error(MSE). The best performing algorithm was a PR model to the 6th degree which produced the best results. It had 98.6% R2, MSE of 14754167.4, MAE of 2191.13 and an RMSE of 3841.12.



Figure - Polynomial Regression Formula

If the data points don’t have a LR than PR can possibly be used but, only if x and y have a relationship. When PR finds the curvilinear relationship between x and y, it then draws a best fit line through the data points. The measurement used to determine the relationship is called r-squared(R2) with value ranges that fall between 0 and 1. The closer the value of x is to 0 means it has no relationship to y, while a value of 1 means x is 100% related to y. In addition to these, anything in the middle means x and y are somewhat related.

A study in 2021 by [32], found it difficult to accurately forecast the trend of COVID-19 active cases too far into the future. The models that were used in this research were LR, Multi-Layer Perceptron (MLP) and Long-Short Term Memory (LSTM). The authors used 78 common features shared by 71 countries. Of these features, 36 were general country related, and the remaining were time series features generated during the pandemic period of 2020 between the 22nd of January to the 20th of July. LR was the better model over a larger time-period, but it’s best R2 value which was greater than 90% came when the forecast was no longer than 3 days with that score starting to drastically fall after 15 days.

LR is used to determine if there is a relationship between the x variable and the y variable by using a straight line of best fit using the least squares method which measures the distance from the data point to the line. In the formula below B0 and B1 are the intercept and slope, respectively.



Figure 5 - Linear Regression Formula

## Classification

Classifications models are used to predict the category of new observations using given labelled data. A classification technique works by learning from a dataset and then organises new data into classes(target variable).

According to [33], who did a comparative analysis of various ML classifications models, on total COVID-19 cases, deaths and recoveries occurring in different states in India, they found that RF performed best at predicting daily COVID cases which produced an accuracy of 83.54%. The various models used in comparison were RF, DT, Multinomial Logistic Regression(MLR), Neural Network(NN) and SVM. The dataset was quite small with 2342 samples with a 70:30 split for training and test data. The timeline of the samples was from the 30th January 2020 to 26th May 2020 with the only features considered being Observation Date, Time, and State/Union Territory.

RF uses the final output from multiple DT by aggregating them together to find the most common result. It utilises bagging and feature randomness thereby, creating multiple DT that are uncorrelated. RF reduces the chances of overfitting, provides flexibility and a way to easily determine important features.

A study conducted by [34], had 3 separate datasets but which had the same samples which were populated with human logistics data that included travel patterns. When the model was run on dataset 1, it could classify its target variable into 5 possible classes for total COVID cases, with the second and third datasets producing 10 and 20 classes respectively. Using dataset 2, its target variable was more complex and difficult to predict than that of dataset 1 conversely, the same can be said about dataset 3 compared to dataset 2. The study used 6 models, K-Nearest Neighbour(KNN), DT, SVM, RF, Time Series Forest Model and LSTM. The data was gathered from the United States, including all 50 states between March 2020 and January 2021 which amounted to 300 days. The 15 features that were used proved to have p-values less than 0.01 indicating a significant statistical relationship with the data with the Pearson Correlation Coefficient backing this up. Some of the features used in the dataset were New Daily COVID-19 Cases, Population Density Data, Current Hospitalizations, Average Temperature, Visits to Fitness Centers, and Visits to Restaurants etc. The model which performed the best on all 3 datasets was the LSTM which gave an F1-score of 0.956320864 for dataset 1, 0.916638327 for dataset 2, and 0.810034019 for dataset 3. As complexity increased from datasets 1-3, it outperformed the other models on every occasion.

LSTM is an extension of an artificial Recurrent Neural Network(RNN) architecture used in Deep Learning(DL). LSTM is equipped to handle time series data because it works by having a different cell state for each layer which allows it to remember information over time.

# **Unsupervised Learning**

[35] used their study to determine the cities in Turkey which needed some form of lockdown measures to stop the spread of COVID. There were 4 risk levels with gradual increasing lockdown measures in Turkey at the time of this study and they were categorised by the number of cases. Risk level 1 had between 0 and 20 cases therefore, no restrictions but, if cases were greater or equal to 100 then it was risk level 4 and required full lockdown. They created 4 datasets based on data released by the Turkish Government which measured COVID cases per 100,000 population in each province between February 20th – 26th 2021. From this data the authors created dataset 1 with Number of Cases as its only feature, dataset 2 with Number of Cases, Population Density as its features, dataset 3 with Number of Cases, Population Density, Average Age, and dataset 4 with Number of Cases, Population Density, Average Age and Air Pollution. These datasets were assembled to determine what variables improved the clustering performance with the only restriction being the data is measured in number of cases per 100,000 population. This study only had clustering models which were Hierarchical Clustering(HC), K-Means Clustering(KMC), Fuzzy C-Means Clustering(FCMC) and Gray Relational Clustering(GRC). Using the Gini Coefficient, Euclidean Distance amongst others, allows clustering models to group samples according to their similarity in data. Measurements techniques used for the internal and external validation were the Silhouette Index(SI), Calinski-Harabasz Index(CHI) and Dunn Index(DI). The closer the value of SI is to 1 demonstrates the analysis of the cluster is high. Additionally, if the CHI and DI has a big value then it indicates the algorithm is performing well. The finding determined GRC was the best model for this research with 0.67 for SI, 362.18 for CHI and 0.046 for DI. Its recommended, that increasing the amount of risk levels and implementing restrictions by regions may increase cluster performance.

# **2.1 Conclusion**

This literature review is to understand the existing state of the art Artificial Intelligence technologies and methods associated with the creation of models that predict COVID-19 case numbers, as well as the data sources to which algorithms are applied. The intent of the review being to identify limitations in the approach currently being taken by the Government of Ireland, so that predictions of COVID-19 cases can be improved. In doing so helping policy makers make better informed public health decisions.

We found that the Irish Epidemiological Model Advisory Group to the National Public Health Emergency Team uses as statistical approach. Fitting their models to patient data from the Computerised Infectious Disease Reporting Database and select, informed assumptions. Predictions are based on nine epidemiological parameters synthesized from a literature review of international best practice among epidemiologists.

B. Tang [36] and Roosa [37] et al applied mathematical models to predict COVID-19 cases and its transmission rate.  Furthermore, to predict COVID-19 instances, several statistical methods such as Autoregressive Integrate Moving Average (ARIMA) were applied by Ceylan [38] and Dehesh [39]. These statistical models were good for predicting the COVID-19 cases for the next 10-15 days but were found to be ineffective over the period of long term. Pratima [40] wrote in her research that Statisical models does not capture the data non-linearity. Furthermore, because these models rely on regression rather than modelling non-linear functions, they are unable to learn the dynamics of an infectious disease's transmission rate. Statistical models like ARIMA, on the other hand, are parametric methods that heavily rely on a number of assumptions. When predicting COVID-19 growth, such theorized factors result in poor data fitting and low accuracy. As a result, statistical models alone are unable to construct generalized models that reflect the current epidemic's randomness.

A potential downside of the IEMAG approach is that patient data is a window into the past, as such is does not consider behaviours happening in the present that may also influence or predict the spread of the virus in the future. Social Media Big Data can provide insights into the actions and intentions of a segment of the population. Google searches for pharmacist, vaccination centre, test centre, pub near me, even anti-vaccine information all provide insights into what people may intent to do (obtain medical treatment, diagnosis, meet with friends etc.). And our review found Social Media content from other platforms has also be found to be predictive of influenza cases. Applying Machine Learning to Big Data offers a way of enhancing the accuracy of the statistical approach and several researchers claim to have made such improvements. Although none of the published papers identified did so for the prediction of case numbers in the Republic of Ireland.

In terms of the different techniques used for ML supervised learning, according to the research tends to perform better than unsupervised learning at predicting COVID cases. Furthermore, it seems that there is more research that uses supervised learning in this space as COVID data is in general, labelled. The best performing models were PR for regression and LSTM for classification.

# **Works Cited**

|  |  |
| --- | --- |
| [1] | S. S. M. Q. L. J. R. G. Y. W. W. Y. L. L. Y. D. G. D. 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] | N. B. N. D. 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. T. 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] | Rialtas na hÉireann, “Irish Epidemiological Modelling Advisory Group to NPHET,” 06 12 2021. [Online]. Available: https://www.gov.ie/en/publication/dc5711-irish-epidemiology-modelling-advisory-group-to-nphet-technical-notes/. [Accessed 07 02 2022]. |
| [6] | IEMAG Epidememiology Modelling subgroup, *Time-Series Based Estimation of the Effective Reproduction Number (Rt),* 05: 21, 2020. |
| [7] | J. Wallinga and P. Teunis, “Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures,” *American Journal of Epidemiology,* vol. 160, no. 6, pp. 509-516, 2004. |
| [8] | University of Malaya Faculty of Medicine, “COVID-19: Understanding R0 and Rt,” Department of Social and Preventative Disease, 21 08 2020. [Online]. Available: https://spm.um.edu.my/2020/09/21/covid-19-understanding-r0-and-rt/. [Accessed 08 02 2022]. |
| [9] | IEMAG Epidemiology Modelling Subgroup, “COVID-19 epidemiological parameters summary document,” Irish Epidemiological Modelling Advisory Group to NEPHET, Dublin, 2020. |
| [10] | IEMAG Epidemiology Modelling Subgroup, “A Hierarchial Model for Monitoring Country Level Incidence,” IEMAG Epidemiology Modelling Subgroup, Dublin, 2020. |
| [11] | NumFOCUS, “About Stan,” [Online]. Available: https://mc-stan.org/. [Accessed 08 02 2022]. |
| [12] | IEMAG Epidemiology Modelling subgroup, “A population-level SEIR model for COVID-19 scenarios,” Irish Epidemiology Modelling Advisory Group, Dublin, 2020. |
| [13] | The R Foundation, “The R Project for Statistical Computing,” The R Foundation , [Online]. Available: https://www.r-project.org/. [Accessed 08 02 2022]. |
| [14] | Oracle, “What is Big Data,” [Online]. Available: https://www.oracle.com/big-data/what-is-big-data/. [Accessed 31 01 2022]. |
| [15] | L. Samaras, E. Garcia-Barriocanal and M.-A. Sicilia, “Comparing Socila media and Google to detect and predict severe epidemics,” Springer Nature, 2020. |
| [16] | 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. |
| [17] | 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. |
| [18] | 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]. |
| [19] | 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. |
| [20] | 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. |
| [21] | 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. |
| [22] | 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. |
| [23] | 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. |
| [24] | 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. |
| [25] | P. Saranya and P. Asha, “Survey on Big Data Analytics in Health Care,” in *Smart Systems and Inventive Technology*, 2019. |
| [26] | 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. |
| [27] | 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. |
| [28] | 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. |
| [29] | 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. |
| [30] | R. Michalski, J. Carbonell and T. Mitchell, “Machine learning: An artificial intelligence approach,” *Artificial Intelligence,* vol. 25, no. 2, pp. 236-238, 1985. |
| [31] | V. Bhadana, A. Jalal and P. Pathak, “A Comparative Study of Machine Learning Models for COVID-19 prediction in India,” *2020 IEEE 4th Conference on Information & Communication Technology (CICT),* pp. 1-7, 2020. |
| [32] | C. Güzeliş, O. Çopur and M. Nakip, “Comparative Study of Forecasting Models for COVID-19 Outbreak in Turkey,” in *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Elazig, 2021. |
| [33] | V. Gupta, A. Gupta, D. Kumar and A. Sardana, “Prediction of COVID-19 Confirmed, Death, and Cured Cases in,” *Big Data Mining and Analytics,* vol. 4, no. 2, p. 116–123, 2021. |
| [34] | A. Li and N. Yadav, “An Adaptable LSTM Network Predicting COVID-19,” in *2021 IEEE International Conference on Digital Health (ICDH)*, Chicago, 2021. |
| [35] | H. Fidan and M. Yuksel, “A comparative study for determining Covid-19 risk levels by unsupervised machine learning methods,” *Expert Systems with Applications,* vol. 190, pp. 1-9, 2022. |
| [36] | X. W. Q. L. N. L. B. S. T. Y. X. B. Tang, Estimation of the transmission risk of the 2019-ncov and its implication for public health interventions, Journal of clinical medicine. |
| [37] | Y. L. R. L. A. K. R. R. J. H. K. Roosa, Real-time forecasts of the covid-19 epidemic in china from february 5th to february 24th 2020, Infectious Disease Modelling. |
| [38] | Z. Ceylan, Estimation of covid-19 prevalence in italy spain and france, Science of The Total Environment. |
| [39] | H. M.-F. a. P. D. T. Dehesh, Forecasting of covid-19 confirmed cases in different countries with arima models. |
| [40] | P. Kumari, Real-time estimation of COVID-19 cases using machine learning and mathematical models - The case of India, Indian Institute of Technology. |
| [41] | C. Keegan and S. Lyons, “Hospital admission probability and length of stay among Covid-19 confirmed cases,” Economic and Social Research Institute, Dublin, 2021. |