Thesis Title

Stochastic Process Models for Short-Term Forecasting of Pandemics

By

Ahmed Qasem Alruwaili
Postgraduate Student

Thesis
Submitted to Flinders University
for the degree of Master of Science (Mathematics)

Master of Science (Mathematics)
College of Science and Engineering
28/11/2022

Supervised by

Dr Greg Falzon
College of Science and Engineering

Acknowledgments

I would like to thank my supervisor Dr Greg Falzon for his assistances and expert advice throughout of this project difficulties. Also, I am grateful to my wife for the love and patient she has always shown to me.

Abstract

The coronavirus disease of 2019 (COVID-19) affected the whole world economically, socially, and healthily. By the fourth quarter of 2022, over 600 million people were infected, with a mortality rate of around 6.5 million. The global response to COVID-19 management has been multi-faceted, involving restrictions, lockdowns, and immunisation programs. Forecasting models have also been widely utilised to estimate future case numbers and inform government policy. Reliable forecasts of disease case numbers are also very important from a medical perspective, as they can significantly assist with resource allocation and planning.

The scientific literature reports on an extensive range of models that have been applied to the modelling and forecasting of the COVID-19 case number dataset. Models investigated range from agent-based computational approaches to statistical stochastic process models and machine learning approaches. Many of the models investigated were successfully applied; however, many were applied in limited country-specific contexts, and substantial limitations were identified regarding the reliability of forecasts. A wealth of data on the COVID-19 pandemic has now been collected, and this data provides an opportunity to address model limitations and develop improved models for future pandemic management.

This project aims to address the gaps identified in the literature by evaluating a wide range of relevant statistical stochastic process models and neural network COVID-19 forecasting models across multiple countries. Furthermore, this project introduces and evaluates a novel modelling approach (ARMA-ELM) that combines both statistical and machine learning models. Model performance was assessed across multiple models and countries, with the ARMA-ELM providing enhanced performance in certain circumstances. Overall, significant differences were found in the COVID-19 data structures between different countries, resulting in no particular model performing best in all circumstances.

Table of Contents

Chapter 1 Introduction

1.1 Introduction

Statistical stochastic process and neural network models have emerged as essential tools in a wide range of domains, including economics, medicine, and biology. These classes of models describe statistical patterns within datasets, accounting for both deterministic and generative random components. Potential uses of statistical stochastic process models include understanding data structure, statistical significance comparisons, prediction and forecasting of future observations. There are several approaches to forecast time series, including classical statistical models and machine learning algorithms. For instance, the ARIMA (Auto-regressive Integrated Moving Average) models are an important class of statistical stochastic process models that are utilised for the modelling univariate time series data. Despite their relatively simple mathematical structure the ARIMA models are well-suited for many applications including COVID-19 incidences (Yang et al. 2020, p. 1417). Major potential benefits of the ARIMA models include the ability to estimate model parameters with a relatively small amount of data and the capability to provide accurate short-term forecasts. The global COVID-19 pandemic has provided an unprecedented amount of data regarding disease transmission dynamics at the population level. Predictive models proved essential to pandemic response, informing health authorities of likely future incidence numbers and hospital admissions. A wide range of models were developed including epidemiological (COVID19 disease transmission dynamics), computational (agent-based models) (Lejenue & Linder 2020), mathematical (differential equations) (Jourdain & Lelievre 2003), statistical (data-driven) (Chintalapudi, Battineni & Amenta 2020), and machine learning (data-driven) (Brunton & Kutz 2022). Data-driven models were particularly successful as they relied on observable patterns rather than implicit assumptions about the mechanisms of disease transmission. The statistical modelling approaches focused on models that were assumed to have particular mathematical structures incorporating a deterministic and random component. The statistical models required the estimation of parameters from the data under the assumed mathematical model structure. In contrast, machine learning methods such as the neural network place less assumptions on the data generative structure and focus instead on attempting to learn a model to describe the data. Both the statistical and machine learning approaches have advantages and disadvantages, the statistical approaches provide a better understanding of the mathematical model generating the data but, in many cases, the assumed model dose not capture all of the complexities associated with the disease transmission process. In contrast, machine learning approaches often provide greater forecast accuracy but at the expense of understanding the structure of the data generation model. It is currently unknown whether the statistical or machine learning approach is best suited to COVID-19 forecasting applications. Furthermore, it might be beneficial to combine statistical and machine learning approaches to obtain the advantages of both paradigms. This project will focus on a kay task faced by forecast modellers during COVID-19 pandemic. Specifically, 7-days ahead daily incidence and mortality forecasts will be estimated and evaluated for a range of state-of-the-art and novel models across five counties. The statistical models evaluated include (ARMA, ARIMA, WARIMA, ARFIMA), neural machine learning models (ANN, ELM), and novel hybrid models (ARMA-ELM). Models will be developed and evaluated using World Health Organisation (WHO) COVID-19 incidences and mortalities datasets at a critical stage in the pandemic (corresponding to the middle of the 4^{th} global wave). By this stage a sufficiently large data set had been collected within each country to allow both statistical and machine learning model fitting. Furthermore, the model training data set will describe a range of complex disease transmission dynamics including the initiation, establishment, and response phases of the pandemic (lockdowns, vaccination campaigns) along with the evolution of COVID-19 variants. This scenario, although limited to one particular point in the pandemic timeline is believed to be highly informative for model comparisons and form the foundation of future research involving more comprehensive investigations across all stages of the pandemic. These future investigations are require considerable computational resources, this research will inform whether such experiments should proceed and guide experimental design.

1.2 Chapters Summary

The introduction and chapters summary have been discussed in Chapter one. In Chapter two the literature review of the statistical stochastic process models and neural network approaches is represented. Chapter three discusses the methodology of the project and introduces the novel ARMA-ELM model. In Chapter four the project results are presented. The discussion of the results is contained in Chapter five. Chapter six provides project conclusion followed by the references list.

Chapter 2 Literature Review

2.1 COVID-19 Forecasting with Stochastic Process Models

2.1.1 Introduction

There have been considerable prior research utilising statistical stochastic process and neural network models to forecast the COVID-19 incidence and mortality numbers. Statistical stochastic process and neural network models are of major scientific interest to this application because of their ability to address random fluctuations in incidence numbers within a mathematical context without full knowledge of the underlying epidemiological processes. The statistical stochastic process and neural network models reviewed are predominately data-driven and only require estimation of model parameters from the COVID-19 incidence and mortality numbers for their implementation. These benefits have encouraged substantial research into a range of potentially suitable models utilising statistical stochastic process and neural networks. As evidenced in this review the stochastic process modelling approach to COVID-19 incidence and mortality numbers are overall quite effective. However, there is considerable scope for scientific research to improve model forecasting performance.

2.1.2 Auto-Regressive Integrated Moving Average (ARIMA) Models

The Auto-Regressive Integrated Moving Average (ARIMA) model is widely applied in demand forecasting and future stock price and electricity prices prediction based on past prices. For instance, Contreras et al. (2003, p. 1018) found that the performance of the ARIMA model needs five hours for the Spanish electricity price market to predict a one-day ahead price forecasting, while the Californian electricity price market needs only two hours for a one-day ahead forecast. Moreover, utilising the ARIMA model, a food manufacturing firm can estimate the demand for its products and make accurate projections ten months in advance (Fattah et al. 2018, p. 7). The ARIMA model is described mathematically via the following form in equation (1).

$$\left(1-\sum_{\mathrm{i}=0}^{\mathrm{p}} \Phi_{\mathrm{i}}\, \beta^{\mathrm{i}}\right)(1-\mathrm{B})^{\mathrm{d}}\, \mathrm{y}_{\mathrm{t}} = \left(1+\sum_{\mathrm{i}=0}^{\mathrm{q}} \theta_{\mathrm{i}}\, \beta^{\mathrm{i}}\right) \varepsilon_{\mathrm{t}}\,. \quad \text{for} \quad \varepsilon_{\mathrm{t}} \sim N(0, \sigma^2) \tag{1}$$

Chakraborty et al. (2022, p. 1035)

where, B is the backshift operator, p and q are ARIMA parameters, and d is the differencing term.

The ARIMA model is also a generalisation of a wider class of models including Auto-Regressive (AR), Moving Average (MA), and Auto-regressive moving Average (ARMA) (the ARIMA model with difference parameter d = 0)

The ARIMA model, as described by Box et al. (2015, p. 11), is an effective model in detecting linear trends in stationary time series data. The ARIMA model has been applied to COVID-19 incidences forecasting in various studies such as Alzahrani, Aljamaan and Al-Fakih (2020, p. 916) used ARIMA model to forecast the daily confirmed incidences of COVID-19 in Saudi Arabia from the 2nd of March 2020 till the 20th of April 2020, also the ARIMA model was utilised to predict a ten-days ahead forecasting for the new confirmed incidences, mortalities and recovery in Pakistan form the 8th of March 2020 to 27th of June 2020 (Khan, Saeed & Ali 2020, p. 1). The ARIMA model was found to be beneficial for COVID-19 forecasting because of the performance and accuracy of the model. According to Ribeiro et al. (2020, p. 7) ARIMA model was the second out of five models that was used to predict COVID-19 confirmed incidences from the beginning of the pandemic until late April 2020 in Brazil. However, the ARIMA model was found to have substantial limitations to COVID-19 new confirmed incidences forecasting, it performed poorly for non-stationary time series datasets. Chakraborty et al. (2022, p. 1058) states that the vast majority of time series datasets relating to epidemics are non-stationary. The ARIMA model is only suitable for stationary time series datasets or time series that can be mathematically transformed to be stationary, hence ARIMA models could be of limited utility for COVID-19 forecasting. Mélard and Pasteels (2000, p. 505) mentioned that the ARIMA model is also limited when it comes to forecasting outliers or extreme values lying outside the general trend captured by the model. These extreme values are very important for the COVID-19 incidence and mortality forecasting. To address these shortcomings, several innovations were proposed, including the Wavelet-Based ARIMA (WARIMA), the Autoregressive Fractionally Integrated Moving Average (ARFIMA), and the Self-exciting Threshold Autoregressive (SETAR) (Chakraborty et al. 2022, pp. 1023-1059 & Chakraborty & Ghosh 2020, pp. 2-9).

2.1.2.1 Wavelet-Based ARIMA (WARIMA) Model

The Wavelet-Based ARIMA (WARIMA) model was designed to address the issue of non-stationary time series dataset (change mean or variance) (Chakraborty et al. 2022, p. 1034). Chakraborty and Ghosh (2020, p. 3) claim that the WARIMA model has the capability to provide information within the signals for both the scale or frequency and time domain. The WARIMA applies the wavelet transform prior to ARIMA modelling. The wavelet transform coefficients are stationary across time and scale under assumptions about the signal properties. The WARIMA model relates signal properties across scales using equation (2).

$$\text{ø}_{m,n}(t) = \frac{1}{\sqrt{|m|}}\,\text{ø}\left(\frac{t-n}{m}\right); \quad m, n \in \mathcal{R} \tag{2}$$

(Chakraborty et al. 2022, p. 1034)

Where, m (≠0) is the scaling parameter, n is the translation parameter. The scaling coefficients $\text{ø}_{m,n}(t)$ thereby describe different properties at different scales (m), allowing the model to describe different trends (e.g., daily, weekly, monthly, etc).

Empirical inspection of the COVID-19 new confirmed incidences and mortalities datasets reveal that it often exhibits non-stationary patterns Figure (2.1) and Figure (2.2).

Figure (2.1)

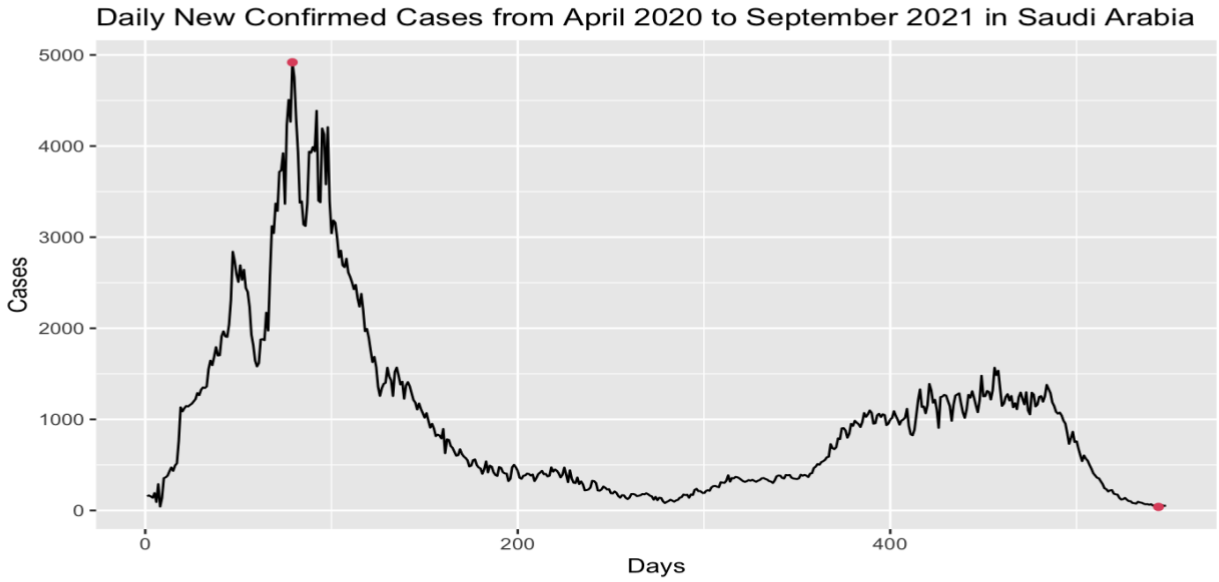
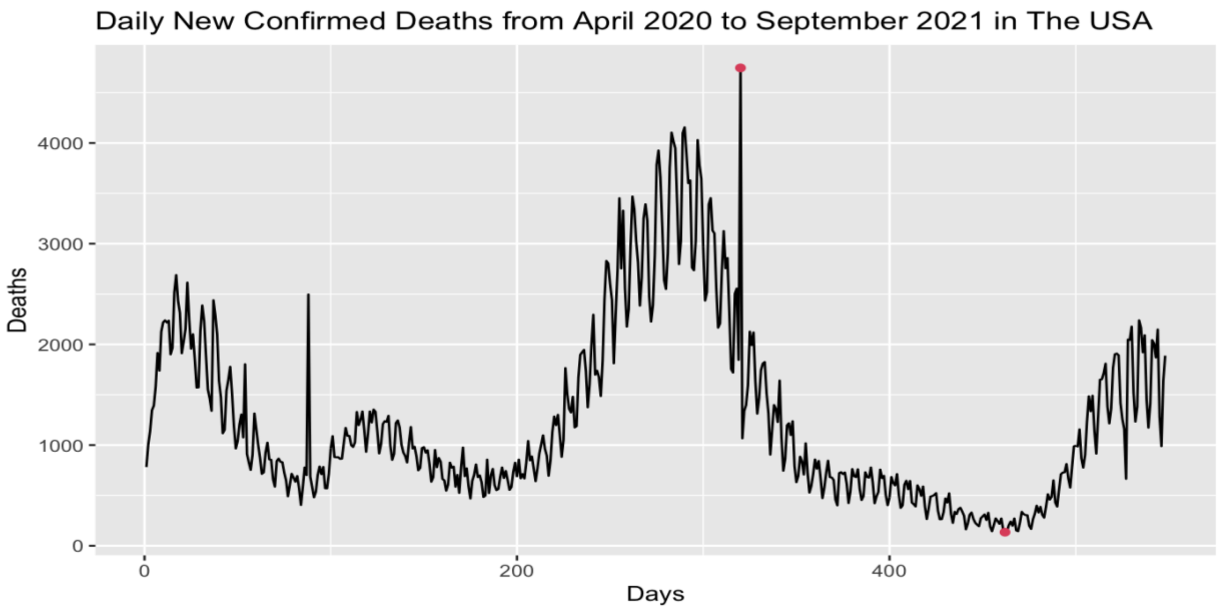


Figure (2.2)



Chakraborty et al. (2022, p. 1034) claim that the WARIMA model performs better in predicting COVID-19 incidences when compared to ARIMA model because most of the COVID-19 time series data are non-stationary. The limitation of the WARIMA model is that it requires large sample of data to describe all of the multi-scale trend (seasonal trends) in order to produce accurate forecasts. This limitation is problematic in the emerging phases of COVID-19 pandemic forecasting, as limited data might be available, and prevent accurate estimation of the WARIMA model in such circumstances.

2.1.2.2 Autoregressive Fractionally Integrated Moving Average (ARFIMA) Model

The Autoregressive Fractionally Integrated Moving Average (ARFIMA) model extends the ARIMA model to better match time series data with long memory (Masa & Diaz 2017, p. 28). As Chakraborty et al. (2022, p. 1035) put it, the ARFIMA models are suitable for time series data that has a slowly long-run mean decay deviation rather than an exponential decay assumed in ARIMA models. Moreover, Liu, Chen & Zhang (2017, p. 2) claim that when it comes to dealing with data that have the Long-range dependency attribute, the ARFIMA model provides a superior fit and outcome when compared to the traditional integer order models in terms of how well it fits the data. The ARFIMA model has the same mathematical form as ARIMA model, however, the difference parameter (d) is restricted to non-negative integers.

$$\left(1-\sum_{i=0}^{p}\Phi_i\,\beta^i\right)(1-B)^d\,y_t=\left(1+\sum_{i=0}^{q}\theta_i\,\beta^i\right)\varepsilon_t\,.\quad \text{for}\quad \varepsilon_t\sim N(0,\sigma^2) \tag{3}$$

Chakraborty et al. (2022, p. 1035)

The ARFIMA model has widely been used in analysing time series such as gold prices, stock returns, trade securities, financial markets, air traffic, and crude oil prices. For example, Armachie (2017, p. 66) found that when compared to the ARIMA models, the ARFIMA model was able to predict values with a reduced standard error as well as a narrower confidence interval in the stock returns. The ARFIMA model is likely very useful for COVID-19 forecasting as such data is likely to have long-range dependencies (Chakraborty et al. 2022, p. 1031).

2.1.2.3 Self-exciting Threshold Autoregressive (SETAR) Model

The Self-exciting Threshold Autoregressive (SETAR) model is an extension of Auto-Regressive models such as ARIMA and ARFIMA. The SETAR model address regime switching behaviour within datasets, which occurs when different magnitudes of response variable correspond to different models. In the SETAR models, each regime is associated with a separate autoregressive (AR) component. According to Davidescu, Apostu & Marin (2021, p. 11) a 2-regime SETAR model may be written mathematically as in equation (4).

$$y_t=\begin{cases}\emptyset_{0(1)}+\sum_{i=1}^{p(1)}\ \emptyset_{i(1)}y_{t-i}+\ \varepsilon_{t(1)}\ ;\ \text{if}\,y_{t-1}\le c\,,\\ \emptyset_{0(2)}+\ \sum_{i=1}^{p(2)}\ \emptyset_{i(2)}y_{t-i}+\ \varepsilon_{t(2)}\ ;\ \text{if}\,y_{t-1}<c\,,\end{cases} \tag{4}$$

Where, $\emptyset_i$ is the coefficients in the regime ε_t is error terms, c is the threshold value at which the regime switches and p is the parameter of AR.

The SETAR model is good at modelling time series data where a higher level of flexibility is required in model parameters (Davidescu, Apostu & Marin 2021, p. 11). According to Tong (1990, p. 321) the SETAR model is applied in predicting future value with assumption that the time series changes the moment the series enters a dissimilar regime. Furthermore, foresight is achieved using the SETAR model on the premise that the time series' behaviour changes when it transitions into a new regime (Chakraborty et al. 2022, p. 1036). The SETAR modelling approach is highly relevant to COVID-19 incidence numbers forecasting, as COVID-19 incidence number datasets often exhibit high non-stationarity and underlying random shifts in incidence numbers driven by unobservable processes. Chakraborty et al. (2022, p. 1052) applied the SETAR model to COVID-19 incidences number forecasting in Brazil, and it performed quite well and out-performed other models such as ARIMA and ARFIMA for short-term forecasting 15 days ahead of COVID-19 incidence numbers. In another words, the SETAR model performs better in terms of accuracy metrics for shorter period forecasting of COVID-19, when compared to other single models. The limitation with this model is that it is less effective if there is no regime switching behaviour. In such scenarios, The SETAR model is over-parametrised and requires greater amounts of data to estimate model parameters.

2.1.3 Neural Network Models

2.1.3.1 Artificial Neural Networks (ANN) Model

The Artificial Neural Networks (ANN) model has been widely utilised in epidemiological forecasting. Applications of the ANN model includes COVID-19, water resources and building electrical energy (Niazkar & Niazkar 2020 p. 4; Pavlicko, Vojteková & Blažeková 2022 p. 1). According to Hilal et al. (2020, p. 4) the mathematical format of the ANN model can be written as listed in equations (5) and (6) below.

$$v = \sum_{i=1}^{m}(wi_{input} * xi) + \beta i_{input} \quad (5)$$

$$y_t = \beta i_{input} + \sum_{i=1}^{n}(vi * wi_{output}) \quad (6)$$

Where, i is the input neurones, m is the first hidden layer, n is the second hidden layer, wi is the node weights and βi is the node bias.

The ANN model calculates the neuron's net input as weighted sum of inputs making it suitable for forecasting of problems of time series dataset (Faraway & Chatfield 1998, p. 234). The ANN model is suitable modelling tool for non-stationary or non-linear datasets, however there are several practical challenges associated with its usage. The selection of the optimal number of model hyper-parameters and hidden layers along with computational complexity make the ANN model more challenging to implement in practice compared to other ARIMA-style models (Maier & Dandy 2000, p. 119). A large number of hidden layers and parameters might also require a greater amount of data to reliably estimate the model. Chakraborty et al. (2022, p. 1050) evaluated ANN models for COVID-19 incidences number forecasting in India. Overall, the ANN model performed the best comparing to the single predicting models for 15 and 30 days ahead forecasts.

2.1.3.2 Autoregressive Neural Network (ARNN) Model

The Autoregressive Neural Network (ARNN) model introduces an auto-regressive component into the ANN model, which provide advantages by incorporating the influence of preceding observations (similar to ARIMA). Jurado et al. (2013, p. 186) claim that due to a decrease in recurrent connections, the autoregressive multi-context recurrent neural network accelerates the training process and is an excellent technique for approximating daily peak demand. According to Chakraborty et al. (2022, p. 1050) the mathematical form of the ARNN model as in equation (7).

$$f(\underline{x}) = c_0 + \sum_{j=1}^{k} w_j \phi\,(a_j + b'_j\ \underline{x}) \quad (7)$$

Where, $\underline{x}$ is p-lagged inputs, c_0, w_j, a_j are connecting strengths, b_j are p-dimensional weight vector and ϕ is a bounded nonlinear sigmoidal function.

The ARNN has been in applications as good at forecasting in arrange of applications, especially for airline and is well-suited for forecasting of non-seasonal time series (Hyndman

& Athanasopoulos 2018, p. 447). Chakraborty et al. (2022, p. 1048) applied the AENN model to the forecasting COVID-19 incidences number in the USA and found that compared to ARIMA, that the ARNN model offers relatively competitive accuracy metrics for 15-day-ahead projections.

2.1.3.3 Extreme Learning Machine (ELM) Model

Recent times have seen a rise in the amount of research utilising neural networks and machine learning methods. The Extreme Learning Machine (ELM) is one such neural machine learning algorithm that has been successfully utilised across a range of domains including illness diagnostics, traffic sign recognition TSR, and image quality (Li & Huang 2022, p. 166). The ELM was developed to train single hidden layer feedforward neural network (SLFN) which is a very popular form of artificial neural network (Wang et al. 2021, p. 2). The ELM has layers of hidden nodes, called neurones, with randomly assigned input weights (Chen 2019, p. 4). The architecture of the Extreme Learning Machine is shown in Figure (2.3).

Figure removed due to copyright restriction.

Figure (2.3) “The architecture of the Extreme Learning Machine”
by (Rajpal et al. 2022, p. 197)

The ELM model performs non-linear regression as per equation (8).

$$y_j = \sum_{i=1}^{L} \beta_i \, g_i \, (a_i \, x_j + b_i) \tag{8}$$

Where, $x_j \in \mathbb{R}^n$ is the training vector, $y_j \in \mathbb{R}^m$ is the target values, a_i is the weight vector between the input layer to the i^{th} hidden node, b_i is the bias, β_i is the weight vector between the i^{th} hidden node and output neurones and $g_i(x_i)$ is a non-linear “activation” function such as the Sigmoid function (9),

$$g_i(x_i) = \left(\frac{1}{1 + e^{(-a_i^T x_i + b_i)}}\right) \tag{9}$$

or Gaussian function (10),

$$g_i(x_i) = e^{(-b_i \|a_i - x_i\|^2)} \tag{10}$$

(Huang, Yu & Gu 2018, p. 109).

The ELM regression process is illustrated in Figure (2.3) the input values $(x_1, \dots, x_j, \dots, x_n)$ are supplied to the ELM network architecture. A set randomly generated weight $(a_1, \dots, a_i, \dots, a_L)$ and biases $(b_1, \dots, b_i, \dots, b_L)$ are then applied to linearly modify the $(x_1, \dots, x_j, \dots, x_n)$ inputs. These linearly modified inputs are then further modified by a per-specified activation function $g(\bullet)$ with additional weights $(\beta_1, \dots, \beta_i, \dots, \beta_L)$ linearly optimised for regression to each observed variable $(y_1, \dots, y_j, \dots, y_m)$.

The ELM model can be described in matrix form as,

$$Y = H\beta \tag{11}$$

where,

$$H = \begin{bmatrix} g_1(a_1 x_1 + b_1) & \cdots & g_1(a_L x_1 + b_L) \\ \vdots & \ddots & \vdots \\ g_1(a_1 x_n + b_1) & \cdots & g_1(a_L x_n + b_L) \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \dots \\ \beta_L \end{bmatrix}^T, \quad Y = \begin{bmatrix} y_1 \\ \dots \\ y_L \end{bmatrix}^T \tag{12}$$

The optimal β weights are selected via minimisation of the least-squares solution.

That is,

$$\|H\hat{\beta} - Y\| = \min_{\beta} \|H\hat{\beta} - Y\| \tag{13}$$

Which is found using the Moore-Penrose generalised inverse of the H matrix.

$$\hat{\beta} = H^{\dagger} Y \tag{14}$$

(Shi, Chen & Li 2018, pp. 1352-1353).

The use of the Moore-Penrose inverse and random weights initialisation provides the ELM with comparatively faster model estimation performance compared to other standard neural network modelling approaches based on back-propagation (Huang, Zhu & Siew 2006, p. 490). Therefore, the ELM seems appropriate for situations requiring rapid prediction and reaction capacity (Huang, Zhu & Siew 2006, p. 499). The ELM models have been used in various real-world applications such as transportation, and animal images where it has demonstrated advantages in term of speed, accuracy, and generalisation (Qing et al. 2020, p. 430). Zhang et al. (2015, p. 4) applied the ELM model on different transportation type datasets such as cars, bikes, and airplanes. They found on their study that the ELM is faster and higher on performance and accuracy compared to other state-of-the-art statistical machine learning models such as the support vector machines (SVM). Khan et al. (2021, p. 1014) applied the ELM model to the forecasting COVID-19 incidence numbers via testing the COVID-19 pneumonia and normal chest computed tomography scans. They found that the ELM classifier outperformed other models such as Quadratic SVM (Q-SVM) and Logistic Regression in terms of predictive accuracy. This application of the ELM was for image datasets, which has a fundamentally different structure compared to the COVID-19 time series datasets that are of interest in this research. Nonetheless, the ELM appears to be a good candidate for further research due to its reported excellence performance of the ELM in COVID-19 incidence number forecasting. Despite these benefits, the ELM model has some limitations that might affect the overall performance. In contrast to other modelling options, such as the ARIMA, the ELM does not include specific components to address the separate seasonal, trend or stochastic components. There is no autoregressive component within the ELM, which instead assumes a linear combination of activation function. In the case of time series data, the Y observations are ordered and likely correlated. Such auto-correlation is not explicitly addressed within the ELM and instead the ELM model is permitted to find an optimal set of β least-squares regression weights. This observation suggests that the ELM is best utilised to describe the overall linear or non-linear structure of the data set, whist the residual component might best be described utilised other models such as ARIMA.

2.1.4 Auto-Correlation Function (ACF)

Auto-correlation function (ACF) is the correlation of a signal with a delayed copy of itself. Autocorrelation, described informally, is the similarity of observations as a function of their distance in time (Flores et al. 2012, p. 3). The ACF describes the various autocorrelations obtained at different lags, r, in the stochastic process. Specifically, the autocorrelation function is defined as:

$$R_{XX}(t_{1,}t_2) = E[X_{t1}\overline{X}_{t2}] \quad (15)$$

(Zięba & Ramza 2011, p. 532)

In practice, on actual discrete time series datasets, the ACF is estimated by

$$R_{XX}(t_1, t_2) = \sum x(t)\overline{x(t-r)} \quad (16)$$

(Zięba & Ramza, 2011, p. 532)

Autocorrelation function can be used to locate missing fundamental frequencies in signals inferred by their harmonic frequencies or to detect periodic signals disguised by noise. It is frequently employed in signal processing when examining functions or collections of values, such as time-domain signals. From a statistical perspective, it is extremely useful to detect hidden dependencies within stochastic processes or identify the order of Auto-regressive or Moving-Average models.

2.1.5 Partial Auto-Correlation Function (PACF)

The partial autocorrelation (PACF) is similar to the ACF; however, it controls the effects of other lags when performing an estimate. It can be considered a conditional autocorrelation; the autocorrelation is calculated conditional on those observations at shorter time lags. The 1st order (lag) partial autocorrelation will be equal to the 1st order autocorrelation (as there are no preceding lags). That is,

$$\emptyset_{11} = corr(x_{t+1}, x_t) = \rho(1) \quad (17)$$

The 2nd order partial autocorrelation is however given by,

$$\emptyset_{22} = corr(x_{t+2} - \hat{x}_{t+2}, x_t - \hat{x}_t) = \frac{Covariance\ (x_t, x_{t-2}|x_{t-1})}{\sqrt{Variance\ (x_t|x_{t-1})}Variance\ (x_t|x_{t-1})} \quad (18)$$

(Dürre, Fried & Liboschik 2015, p. 207)

More generally,

$$\emptyset_{hh} = corr(x_{t+h} - \hat{x}_{t+h}, x_t - \hat{x}_t), \quad h \geq 2 \quad (19)$$

The PACF can be used to examine the autocorrelation structure of a process with the effects of the other lags removed. A useful aspect of the PACF is that it can be utilised to identify order of AR process models. The theoretical PACF for an AR model "shuts off" after the model's order. The term "shuts off" refers to the theoretical limit beyond which the partial autocorrelations are equal to 0. In other words, the number of partial autocorrelations that are non-zero determines the order of the AR model.

Chapter 3 Methodology

3.1 Overview: The COVID-19 Pandemic:

Pandemics are large disease outbreaks that affect several countries and pose major health, social, and economic risks (Madhav et al. 2018, p. 315). A quick-moving pathogen spreading across the globe has the potential to kill tens of millions of people, disrupt economies, and destabilise national security just as the Spanish flu influenza, HIV/AIDS, and COVID-19 has demonstrated. On the 31st of December 2019, a group of people who worked at the Huanan Seafood Open Market in Wuhan, Hubei Province, all got infection pneumonia infection (World Health Organisation 2020). In the early stages of the COVID-19, some of the patients had been in contact with a wholesale seafood market and this suggests animal to person transmission (Yang et al. 2020, p. 2). A total of 1975 people have been infected and 25 reported mortalities due to the COVID-19 by the end of January 2020 in mainland China (Wang, Tang & Wei 2020, p. 443). The first confirmed incidences reported outside mainland China was a person who travelled to Thailand form Wuhan on the 8th of January 2020 (World Health Organisation 2020). From this date, COVID-19 pandemic spread rapidly across the globe. The COVID-19 outbreak was designated a worldwide pandemic by the World Health Organization (WHO) on the 11th of March 2020 (Cucinotta & Vanelli 2020). According to the WHO by 2022 the total number of people who contracted COVID-19 is over 600 million with over 6.5 million mortalities (https://covid19.who.int).

3.2 Datasets:

The COVID-19 daily prevalence datasets were collected from the World Health Organisation's (WHO) official website (https://covid19.who.int). The WHO COVID-19 website is an authoritative data source with information reported by countries across the world. Table (3.1) summarises the data collected from the WHO COVID-19 website across countries and data collection periods. The daily new confirmed incidences and mortalities datasets were collected for five countries including Brazil, India, Saudi Arabia, Spain, and the United States of America (USA) between April 2020 and September 2021. These countries were selected due to the disproportionally high impact of the COVID-19 pandemic in these countries compared to neighbourhood countries. For instance, Saudi Arabia had the highest new confirmed incidences during this period comparing to other Gulf Arab countries. Moreover, Brazil and the USA also had the highest number of COVID-19 incidences within the South and North American regions. The specific time frames collected were because of the difficulties the counties faced due to not understanding the virus, multiple waves, and variants such as alpha and delta.

Table (3.1). COVID -19 Incidences and Mortalities Datasets Summary

	Location	Total	Min	Max	Date	Length
Daily Confirmed incidences	Brazil	21283107	0	150106	From April 2020 To September 2021	548 DAYS
	India	33664783	336	414188		
	Saudi Arabia	545527	39	4919		
	Spain	4814258	113	40902		
	The United States of America	42809418	8363	294541		
Daily Confirmed Mortalities	Brazil	594740	23	4249	From April 2020 To September 2021	548 DAYS
	India	451490	5	6148		
	Saudi Arabia	8693	0	58		
	Spain	85837	0	913		
	The United States of America	685243	135	4746		

3.3 ARMA-ELM Model:

The literature review demonstrated that statistical time series models such as ARIMA, WARIMA, AFRIMA, and non-linear models such as ANN, and ELM are helpful for modelling and predicting time series similar to the COVID-19 datasets. A novel approach might be to combine the benefits of statistical and machine learning models. The statistical models would provide partial understanding of the data generative process whilst the machine learning models could describe the non-linear components.

The combined models are the Extreme Learning Machine (ELM) which is a type of feedforward artificial neural network that is trained using a single step learning algorithm, in contrast to the iterative learning algorithms used in other types of neural networks. And the Autoregressive Moving Average (ARMA) model is a type of statistical model that is used to describe the temporal dependencies in time series data.

It is possible to combine an ELM model with an ARMA model by using the ELM model to predict the next value in the time series based on the previous values, which can be modelled using an ARMA model. To do this, the input to the ELM model would be the past values of the time series, and the output would be the next value in the series. The ELM model could then be trained to minimise the difference between the predicted next value and the actual next value in the time series.

Alternatively, the ELM model could be used to predict the residuals of an ARMA model, which are the differences between the observed values of the time series and the values predicted by the ARMA model. This approach could potentially improve the accuracy of the ARMA model by capturing any remaining temporal dependencies that are not captured by the ARMA model. The mathematical form of the combined model is displayed as follow in equation (20).

$$\begin{aligned} y_t &= x_t + \varepsilon_t, \qquad \text{for} \quad \varepsilon_t \sim N(0, \sigma^2) \\ &= d_t + s_t + \varepsilon_t \end{aligned} \tag{20}$$

Where, y_t is the observation of time series dataset, ε_t is the error term, d_t is the ELM model,

$$d_t = \sum_{i=1}^{L} \beta_i \, g_i \, (a_i \, x_t + b_i) \tag{21}$$

s_t is the ARMA model,

$$s_t = y_t = \alpha_0 + \sum_{i=0}^{p} \beta_i \, y_{t-i} + \gamma_t - \sum_{j=0}^{q} \alpha_j \, \gamma_{t-j} \tag{22}$$

where, $\gamma_t \sim N(0, \sigma^2)$ is the error term.

Combining two or more time series models could improve the quality and performance of these models. The benefit of combining models is that some models have difficulty dealing with non-stationary such as ARMA time series data sets. In contrast, others, such as ELM, have no problem dealing with stationary and non-stationary time series datasets. The Auto-Regressive Moving Average (ARMA) and ELM algorithms have been discussed earlier in the literature review section. Furthermore, ARMA and ELM algorithms could perform much better with COVID-19 time series datasets for short-term forecasting. Several models such as the SETAR and ARNN were reviewed but not evaluated dur to difficulties in practical implementation within the timelines available in the project.

3.4 Experimental Analysis:

3.4.1 Model Performance Metrics:

When forecasting time series models, the different methods of the model performance metrics needed to evaluate the performance of these models. The most well-known and widely used model performance metrics are as follow:

- Root Mean Square Error (RMSE)

RMSE is a measure of the difference between predicted and actual values that is popular because it is in the same unit as the original data, which makes it easy to interpret. It is calculated as the square root of the mean squared error (MSA), which is the mean of the squared differences between the predicted and actual values.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (23)$$

(Chai & Draxler 2014, p. 1248)

- Mean Absolute Error (MAE)

MAE is another measure of the difference between predicted and actual values, but it is less sensitive to outliers than RMSE. It is calculated as the mean of the absolute differences between the predicted and actual values.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (24)$$

(Chai & Draxler 2014, p. 1248)

- Mean Absolute Percentage Error (MAPE)

MAPE, is a measure of the difference between the predictions made by a model and the true values, expressed as a percentage. It is calculated by taking the average of the absolute differences between the predictions and the true values, divided by the true values, and multiplied by 100.

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \quad (25)$$

(Chakraborty et al. 2022, p. 1045)

- Symmetric Mean Absolute Percentage Error (SMAPE)

SMAPE, is a variant of MAPE that is more suitable for use when the true values can be negative. It is calculated by taking the average of the absolute differences between the predictions and the true values, divided by the average of the true values and the absolute value of the predictions, and multiplied by 100.

$$\text{SMAPE} = \frac{100}{n} \sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|}{(|\hat{y}_i| + |y_i|)/2} \qquad (26)$$

(Chakraborty et al. 2022, p. 1045)

Where, y_i are actual values of the time series response and $\hat{y}_i$ are the forecasts of the time series y_i response several (n) time series in the future. However, the forecasting model that performs the best is the one that has the most accurate metrics as assessed on data previously unseen (not used to fit the model) but several (n) time steps in the future. A 7-days ahead forecast was evaluated in this project as it corresponds to a weekly response strategy. Note that these metrics are also summative over this time period, all daily forecast and a 7-day ahead forecast. Model performance metrics may be useful for measuring the average prediction error and tracking forecast performance over time. Understanding model forecast performance can be utilised to guide the selection of the most appropriate model.

3.4.2 Methods:

There are several methods used to analysis the COVID-19 confirmed incidences and mortalities time series datasets such as ARMA, ARIMA, ELM and ELM-ARMA. However, these methods respond differently regarding to the performance of the model performance metrics such as RMSE, MAE, MAPE, and SMAPE. These metrics were selected and utilised due to their wide use within the scientific literature for evaluation of forecast model performance. Table (3.2) is showing the models that used for specific COVID-19 time series datasets and the model performance metrics. Table (3.3) displays the functions and packages that used in R programming for each model. Table(3.4) shows the functions used in Microsoft Excel software to calculate the model performance metrics used in this project.

Table (3.2) Methods, Data and Model Performance Metric

Methods	Data	Model Performance Metrics
ARMA	COVID-19 Incidences and Mortalities Datasets for Five Countries	RMSE MAE MAPE SMAPE
ARIMA		
WARIMA		
ARFIMA		
ANN		
ELM		
ARMA-ELM		

Table (3.3) The Implementation of the Functions and Package used in R programming

Models	Functions	Packages	Resources
ARMA	arima	forecast	(Hyndman & Khandakar 2008)
ARIMA	auto.arima	forecast	(Hyndman & Khandakar 2008)
WARIMA	WaveletFittingarma	WaveletArima	(Paul & Samanta 2017)
ARFIMA	arfima	forecast	(Hyndman & Khandakar 2008)
ANN	mlp	nnfor	(Kourentzes 2022)
ELM	elm	nnfor	(Kourentzes 2022)
ARMA-ELM	----	----	----

Table (3.4) The Implementation of the Functions used in Excel

Model Performance Metrics	Functions
RMSE	=sum() =min() =max()
MAE	
MAPE	
SMAPE	

3.5 Summary:

Accurate forecasting models can help with the management of the COVID-19 pandemic. A large amount of data has now been collected that can assist with the development and refinement of existing models. The COVID-19 confirmed incidences and mortalities time series datasets for five countries such as (Brazil, India, Saudi Arabia, Spain, and the USA) were collected daily from April 2020 to September 2021. These data sets were gathered to forecast using different statistical stochastic process and neural network models such as (ARMA, ARIMA, WARIMA, WARFIMA, ANN, ELM, and ARMA-ELM). These models have different performance which can be checked via virous model performance metrics such as (RMSE, MAE, MAPE, and SMAPE). Assessment of model performance across multiple countries through these metrics will be used to gain further insights regarding the most suitable model for COVID-19 pandemic forecasting.

Chapter 4 Results

4.1 Analysis of Results:

In this chapter, several stochastic process time series models are used to perform forecasts of confirmed incidences and mortalities of COVID-19 daily datasets. The COVID-19 time series datasets divided into two periods, (i) the training dataset and (ii) testing dataset. For the daily new confirmed incidences and mortalities, the first 541 days of the time series used for training, while 7 days ahead used for forecasting test set. Fitting a model using training data and testing it with test data is typical. Most test data comparisons involve distinct forecast timeframes. Our forecasting model was calculated on training data and tested on the latest seven observations. Forecast error will be measured summatively 7-days ahead. Tables (4.1) and (4.2) show the training (shown in black) and testing (shown in red) datasets for Brazil, India, Saudi Arabia, Spain, and the USA. These tables also present the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots for their respective time series datasets. The ACF and PACF provide insights into the different statistic dependency structures in the COVID-19 datasets. For the purpose of short-term forecasting of COVID-19 confirmed incidences and mortalities in five countries, seven distinct statistical stochastic process and shallow neural network forecasting models (ARMA, ARIMA, WARIMA, ARFIMA, ANN, ELM, and ARMA-ELM) have been evaluated as potential competitors. Forecasts for the next 7 days have been created for each model, and model forecasting metrics have been calculated, in order to discover which models, provide the most accurate forecasts. The performance of each model was evaluated using several error metrics (RMSE, MAS, MAPE, and SMAPE), and the models ranked according to optimal performance on each of these metrics for forecasting COVID-19 according to each country's datasets.

Table (4.1) COVID-19 time series of daily confirmed incidences and corresponding ACF and PACF plots

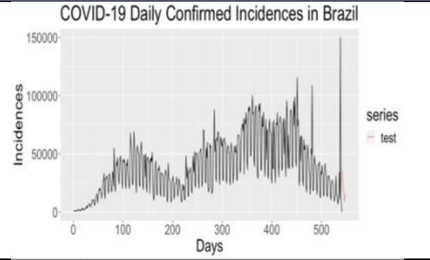
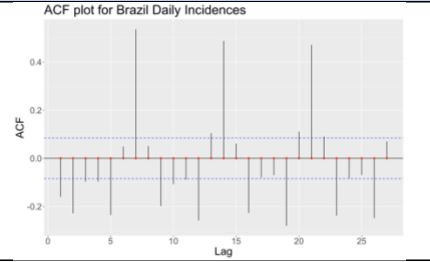
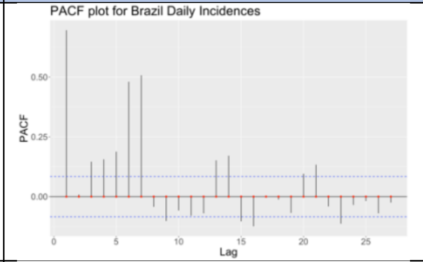
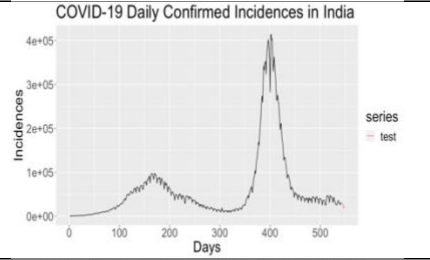
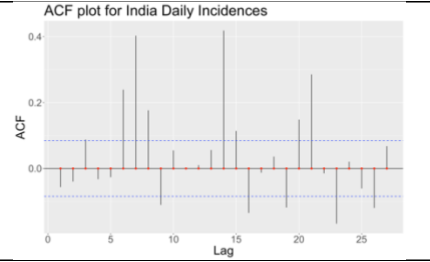
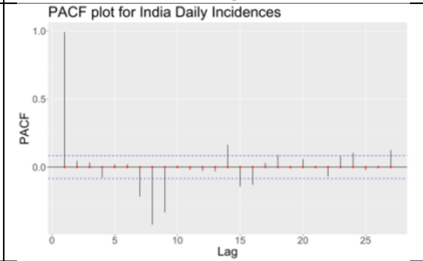
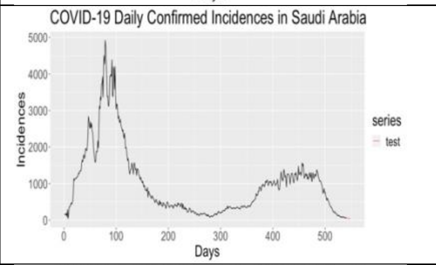
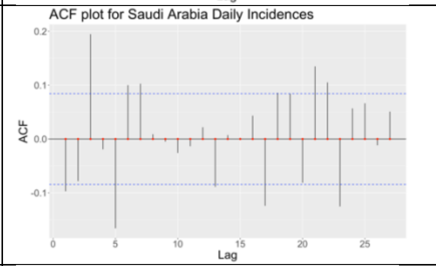
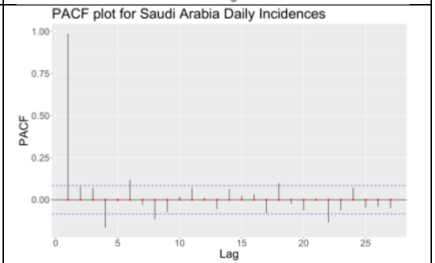
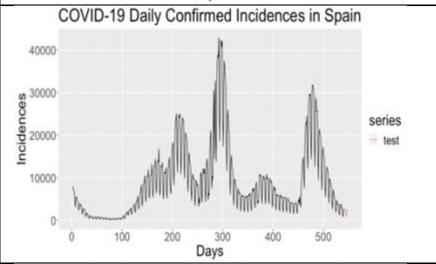
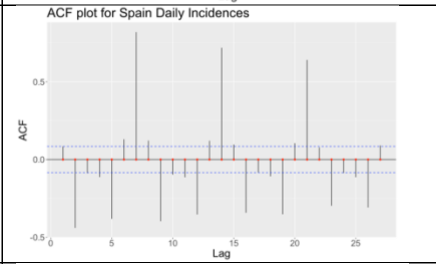
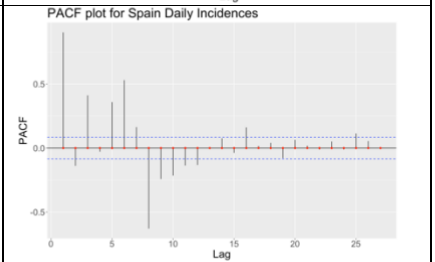
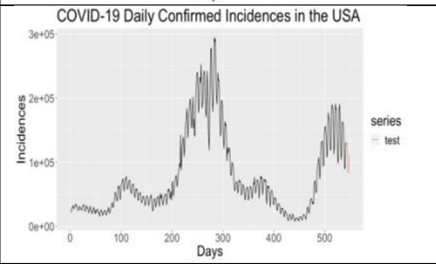
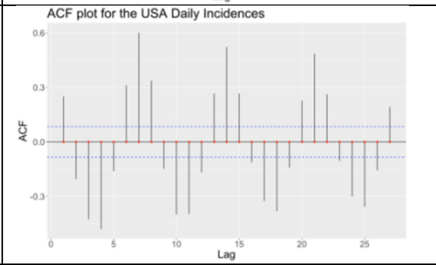
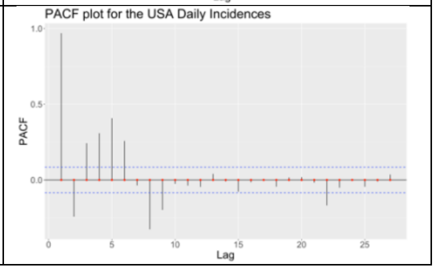
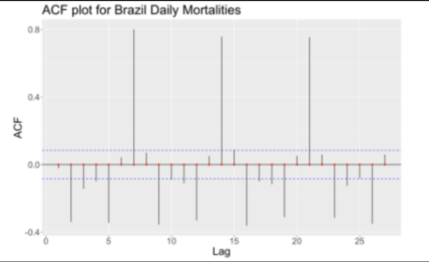
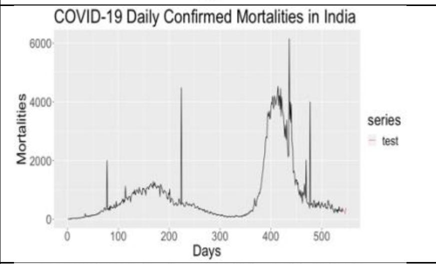
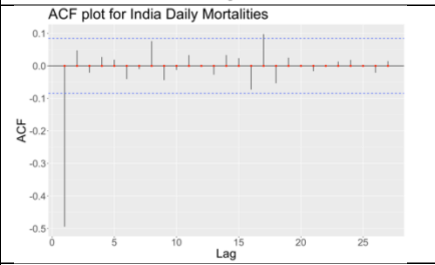
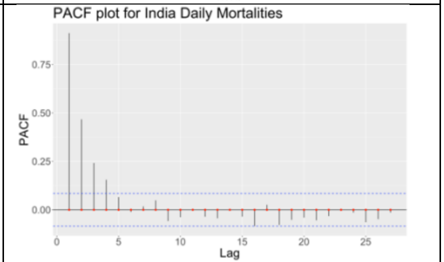
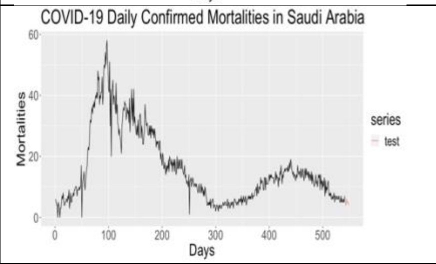
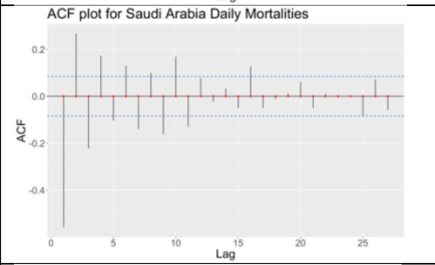
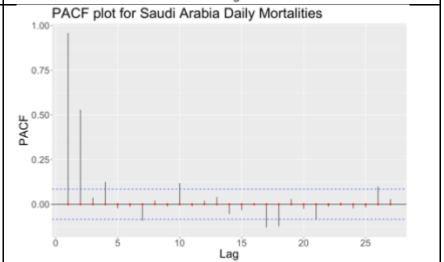
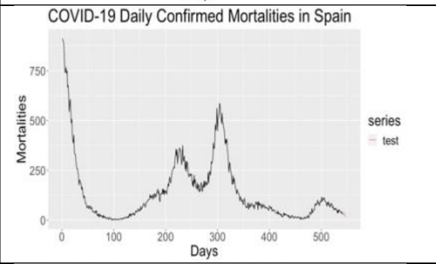
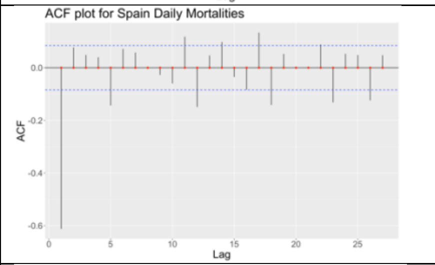
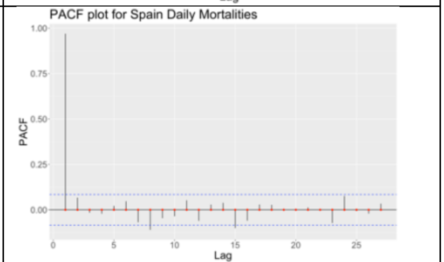
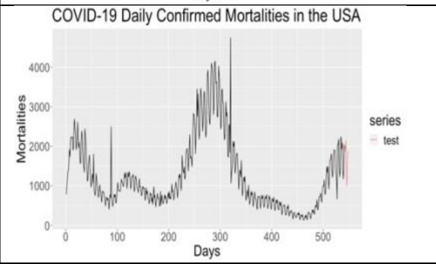
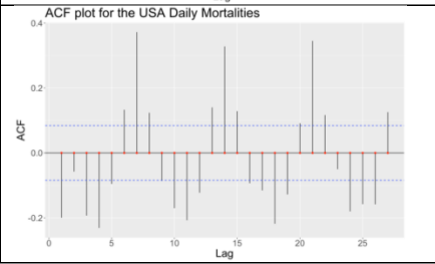
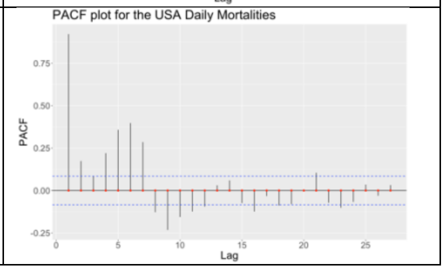
Location	Time Series	ACF Plot	PACF Plot
Brazil			
India			
Saudi Arabia			
Spain			
USA			

Table (4.2) COVID-19 time series of daily confirmed mortalities and corresponding ACF and PACF plots

Location	Time Series	ACF Plot	PACF Plot
Brazil			
India			
Saudi Arabia			
Spain			
USA			

4.1.1 Results for Brazil COVID-19 Datasets:

4.1.1.1 Daily Confirmed Incidences:

The model performance metrics for 7-days ahead forecasts for the COVID-19 daily confirmed incidences in Brazil are shown in table (4.3). The table displays that the novel model ARMA-ELM has the best performance based on the model performance metrics results for 7-days ahead forecasts. Also, the table shows competitive model performance measures between the statistical stochastic process models ARMA and WARIMA and shallow neural network model ANN and ELM respectively.

Table (4.3) Results of Model Performance Metrics for COVID-19 Daily Incidences in Brazil

Model	Rank	7-Days Ahead Forecast			
		RMSE	MAE	MAPE	SMAPE
ARMA	2	17468.24	15805.13	103.8758	0.6907718
ARIMA(3,1,2)	7	30929.31	26272.625	168.3126	0.7927868
WARIMA	5	18358.17	14140.16	79.89341	0.4975719
ARFIMA(5,0,1)	6	27305.34	22286.41	126.0224	0.84137
ANN	3	17428.01	16977.48	113.5309	0.6616902
ELM	4	18233.9	16364.62	119.5199	0.6489345
ARMA-ELM	1	8949.213	7775.971	51.13834	0.3922376

4.1.1.2 Daily Confirmed Mortalities:

ARIMA(3,1,2) model performs better in terms of model performance measures for a 7-days ahead forecast in case of Brazil COVID-19 daily confirmed mortalities dataset than other single models. The WARIMA model also indicated competitive model performance metrics for Brazil COVID-19 daily confirmed mortalities data for 7-days ahead forecasts.

Table (4.5) Results of Model Performance Metrics for COVID-19 Daily Mortalities in Brazil

Model	Rank	7-Days Ahead Forecast			
		RMSE	MAE	MAPE	SMAPE
ARMA	6	447.7734	370.6344	118.7944	0.5594492
ARIMA(3,1,2)	1	160.2656	133.2710	31.19412	0.2735234
WARIMA	2	205.9289	169.3075	35.83052	0.3541215
ARFIMA(3,0,3)	5	377.6871	293.0789	96.97163	0.4874408
ANN	7	602.4776	567.4116	110.7475	1.66141
ELM	3	303.7401	268.1312	268.1312	0.5407471
ARMA-ELM	4	341.9699	304.4522	52.20345	0.6264325

4.1.2 Results for India COVID-19 Datasets:

4.1.2.1 Daily Confirmed Incidences:

Table (4.6) shows that overall, the ARMA-ELM model has the best model performance whilst the ARFIMA(1,0.3,1) has the best performance of all statistical stochastic process time series models. In contrast, the WARIMA was the worst model for 7-days ahead forecast because it has the largest error metrics performance. On the other hand, the shallow neural network models (ANN, ELM) performed similarly based on the model performance metrics for 7-days ahead forecasts.

Table (4.6) Results of Model Performance Metrics for COVID-19 Daily Incidences in India

Model	Rank	7-Days Ahead Forecast			
		RMSE	MAE	MAPE	SMAPE
ARMA	6	8602.309	7053.537	33.0063	0.2603236
ARIMA(0,1,0)	3	8162.070	6700.286	31.356635	0.2498328
WARIMA	7	14294.09	13320.25	58.86671	0.4301177
ARFIMA(1,0.3,1)	2	7579.647	6006.097	28.47378	0.2283468
ANN	5	8534.936	6960.871	32.63198	0.2575118
ELM	4	8409.729	6933.597	32.39021	0.2568567
ARMA-ELM	1	4826.279	3777.928	17.78628	0.1544262

4.1.1.2 Daily Confirmed Mortalities:

For 7-days ahead forecasts, table (4.7) clarify that the ARIMA(1,1,2) is found to have the best scores based on the model performance metrics. The ARMA-ELM, ELM, and WARIMA models have competitive model performance metrics in the same forecasting test period. From table (4.7) ANN model has the largest error metrics for 7-days ahead forecast. This means that ANN model is the worst model among statistical stochastic process and shallow neural network models.

Table (4.7) Results of Model Performance Metrics for COVID-19 Daily Mortalities in India

Model	Rank	7-Days Ahead Forecast			
		RMSE	MAE	MAPE	SMAPE
ARMA	6	70.23672	53.71193	23.17055	0.1876006
ARIMA(1,1,2)	1	58.00556	43.60589	17.95554	0.1572433
WARIMA	3	62.01478	47.30692	18.76769	0.1704964
ARFIMA(2,0.26,2)	5	69.47171	53.08754	22.89292	0.1857575
ANN	7	120.4137	100.3295	41.46147	0.3078169
ELM	4	62.09704	46.94246	20.02149	0.1674212
ARMA-ELM	2	60.69611	44.51769	18.88216	0.1598831

4.1.2 Results for Saudi Arabia COVID-19 Datasets:

4.1.2.1 Daily Confirmed Incidences:

Among the single models shown in table (4.8), the ARIMA(4,1,2) model performs the best for 7-days ahead forecasts based on the model performance metrics results. In comparison, the ARMA and ARFIMA(1,0.08,1) models have competitive model performance metrics results for 7-days ahead forecasts with each other.

Table (4.8) Results of Model Performance Metrics for COVID-19 Daily Incidences in Saudi Arabia

Model	Rank	7-Days Ahead Forecast			
		RMSE	MAE	MAPE	SMAPE
ARMA	3	35.53531	32.6039	66.85814	0.4785586
ARIMA(4,1,2)	1	5.905277	5.00311	10.860020	0.1014873
WARIMA	2	21.66	15.35638	29.2011	0.4061347
ARFIMA(1,0.08,1)	4	38.86632	35.55259	72.63117	0.5079171
ANN	6	80.81619	75.30142	152.2493	0.8233498
ELM	7	255.537	234.2958	468.5573	1.325971
ARMA-ELM	5	53.42374	52.39692	107.4632	0.6830167

4.1.2.2 Daily Confirmed Mortalities:

Table (4.9) shows a strong competition between most of the single models in term of the model performance metrics for 7-days ahead forecasts. However, the WARIMA model has the lowest model performance metrics scores which means that the performance of the model is the best compared to other models such as ARMA, ARIMA(1,1,2) and ANN.

Table (4.9) Results of Model Performance Metrics for COVID-19 Daily Mortalities in Saudi Arabia

Model	Rank	7-Days Ahead Forecast			
		RMSE	MAE	MAPE	SMAPE
ARMA	3	1.623874	1.481426	32.8618	0.2718467
ARIMA(1,1,2)	2	1.452730	1.269477	28.51576	0.2382759
WARIMA	1	1.165185	1.00046	22.49156	0.1936144
ARFIMA(2,0,0)	5	2.074942	1.904782	42.03926	0.3331312
ANN	4	1.656425	1.469885	32.84141	0.2694078
ELM	7	3.671586	3.484829	75.34023	0.5274326
ARMA-ELM	6	2.289793	2.196469	47.80663	0.3755856

4.1.4 Results for Spain COVID-19 Datasets:

4.1.4.1 Daily Confirmed Incidences:

Table (4.10) shows the model performance metrics scores for the COVID-19 daily confirmed incidences forecasting in Spain. ARIMA(5,1,2) is found to have the best performance in terms of the model performance metrics for 7-days ahead forecasts among statistical stochastic process and shallow neural network models for COVID-19 daily confirmed incidences in Spain.

Table (4.10) Results of Model Performance Metrics for COVID-19 Daily Incidences in Spain

Model	Rank	7-Days Ahead Forecast			
		RMSE	MAE	MAPE	SMAPE
ARMA	7	4139.345	3789.202	255.5722	0.9467358
ARIMA(5,1,2)	1	317.989	270.7406	17.97392	0.155588
WARIMA	2	439.97	310.1688	21.70497	0.1769777
ARFIMA(0,0.49,2)	6	2785.83	2582.656	177.4441	0.7940297
ANN	4	793.1334	476.3264	46.77364	0.2674402
ELM	5	1649.367	1368.386	104.2062	0.5330315
ARMA-ELM	3	663.5095	642.337	41.15941	0.3612768

4.1.4.2 Daily Confirmed Mortalities:

The COVID-19 daily confirmed mortalities forecasting models in Spain have different model performance metrics results. Table (4.11) shows that ARMA-ELM performs the best in terms the model performance metrics for 7-days ahead forecasts. Additionally, the table also shows competitive model performance metrics results for 7-days ahead forecast among the single models such as ARMA, ARIMA(0,2,4), and ARFIMA(2,0.06,3) models.

Table (4.11) Results of Model Performance Metrics for COVID-19 Daily Mortalities in Spain

Model	Rank	7-Days Ahead Forecast			
		RMSE	MAE	MAPE	SMAPE
ARMA	6	14.29874	12.79646	59.91859	0.4286964
ARIMA(0,2,4)	4	13.79661	12.50071	58.06993	0.4218927
WARIMA	7	19.39089	16.70897	79.58947	0.509387
ARFIMA(2,0.06,3)	5	13.97416	12.66925	58.84455	0.4260493
ANN	3	11.36071	10.23447	47.63107	0.3628892
ELM	2	9.49796	8.90565	40.33561	0.3239295
ARMA-ELM	1	7.267706	6.804329	30.67781	0.2605935

4.1.3 Results for the USA COVID-19 Datasets:

4.1.3.1 Daily Confirmed Incidences:

Forecasting models of the daily incidences in the USA shown in table (4.12) illustrates that the novel model ARMA-ELM has the best performance based on the model performance metrics results for 7-days ahead forecasts. However, WARIMA and ARFIMA(5,0.13,2) models also have competitive model performance metrics.

Table (4.12) Results of Model Performance Metrics for COVID-19 Daily Incidences in USA

Model	Rank	7-Days Ahead Forecast			
		RMSE	MAE	MAPE	SMAPE
ARMA	5	27731.59	22869.58	23.9978	0.2010715
ARIMA(0,1,5)	7	39683.07	37035.528	36.96837	0.2999612
WARIMA	2	18721.4	17887.17	16.98363	0.1549072
ARFIMA(5,0.13,2)	3	19156.81	17157.44	14.96898	0.1380828
ANN	6	28763.36	26431.98	26.40542	0.2251015
ELM	4	22481.25	19642.13	20.07649	0.1754591
ARMA-ELM	1	17566.55	14498.01	15.01234	0.135726

4.1.3.2 Daily Confirmed Mortalities:

Results shown in table (4.13) clarify that among all the statistical stochastic process and shallow neural network models that used in this project, ARIMA(3,1,2) performs the best in terms of the model performance metrics for 7-days ahead forecasts. Further, ANN, ARFIMA(1,0.49,0), and ELM models also have competitive model performance metrics for the same predicting period.

Table (4.13) Results of Model Performance Metrics for COVID-19 Daily Mortalities in USA

Model	Rank	7-Days Ahead Forecast			
		RMSE	MAE	MAPE	SMAPE
ARMA	5	383.0114	328.3725	22.97733	0.2060458
ARIMA(3,1,2)	1	261.1251	195.0309	14.07068	0.1274668
WARIMA	7	466.0928	364.8011	21.09943	0.2392982
ARFIMA(1,0.49,0)	3	352.5501	303.0633	20.97745	0.1928247
ANN	2	305.1653	244.9936	18.46981	0.1563767
ELM	4	361.2123	293.1694	22.06462	0.1838653
ARMA-ELM	6	388.5118	336.739	23.02036	0.2111565

4.2 Models Summary

The project used seven different time series models to forecast the COVID-19 incidences and mortalities for the five countries. Figure (4.1) displays the number of times each models had top performance (according to the model performance metrics) for COVID-19 incidences datasets. These results indicate that for the COVID-19 incidence data showed that ARIMA model performs better for both Saudi Arabia and Spain time series. In contrast, the novel model ARMA-ELM performs better for Brazil, India, and the USA time series datasets.

Figure (4.1) Top Model Counts for COVID-19 Incidences datasets across all countries

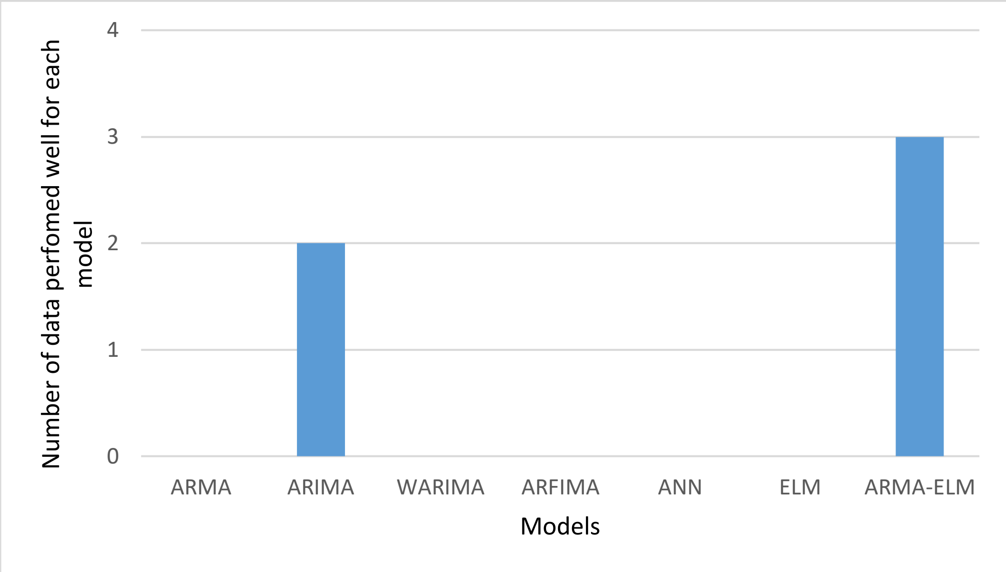
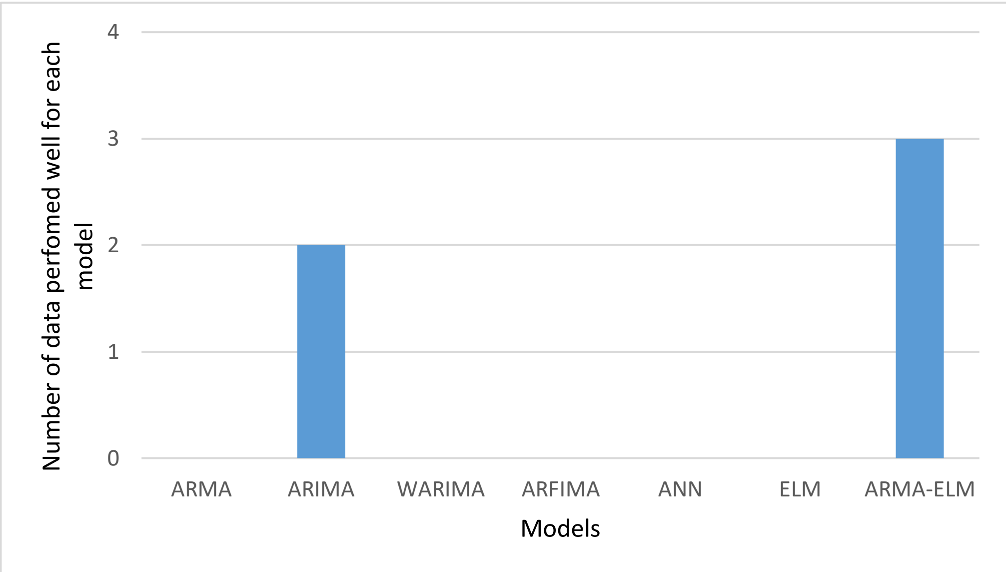


Figure (4.2) displays the top-model performing counts for the COVID-19 mortalities datasets according to the model performance metrics. The ARIMA model performed the best three times (for the COVID-19 mortalities datasets of Brazil, India, and the USA). Whereas the WARIMA and ARMA-ELM models both performed the best once (for the COVID-19 mortalities datasets of Saudi Arabia and Spain respectively).

Figure (4.2) Top Model Counts for COVID-19 Mortalities datasets across all countries



Figures (4.3), (4.4), (4.5), and (4.6), respectively illustrate the mean, minimum, and maximum of the RMSE, MAE, MAPE, and SMAPE metrics for the COVID-19 incidences datasets across all countries.

Figure (4.3) RMSE for Incidences datasets for all countries

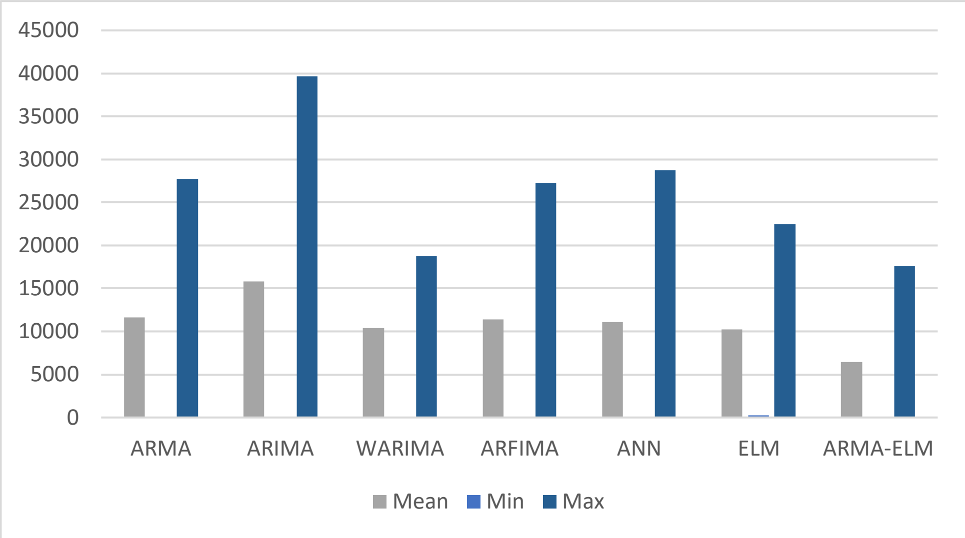


Figure (4.4) MAE for Incidences datasets for all countries

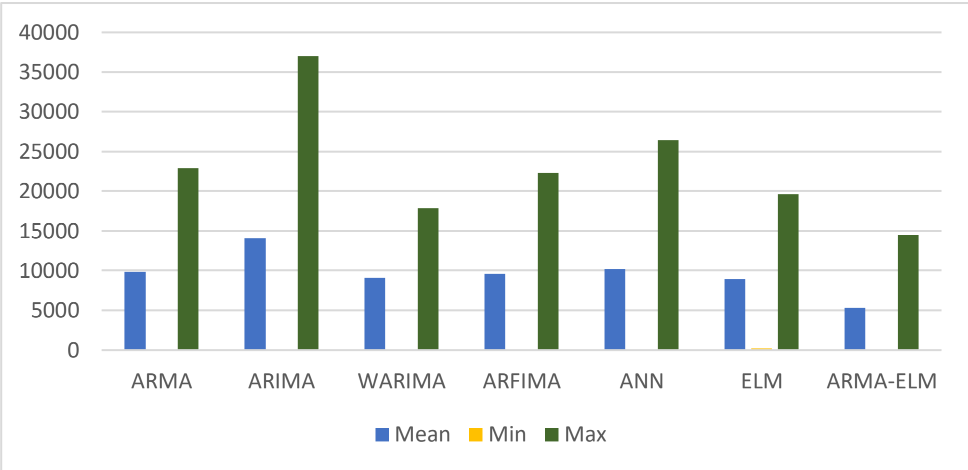


Figure (4.5) MAPE for Incidences datasets for all countries

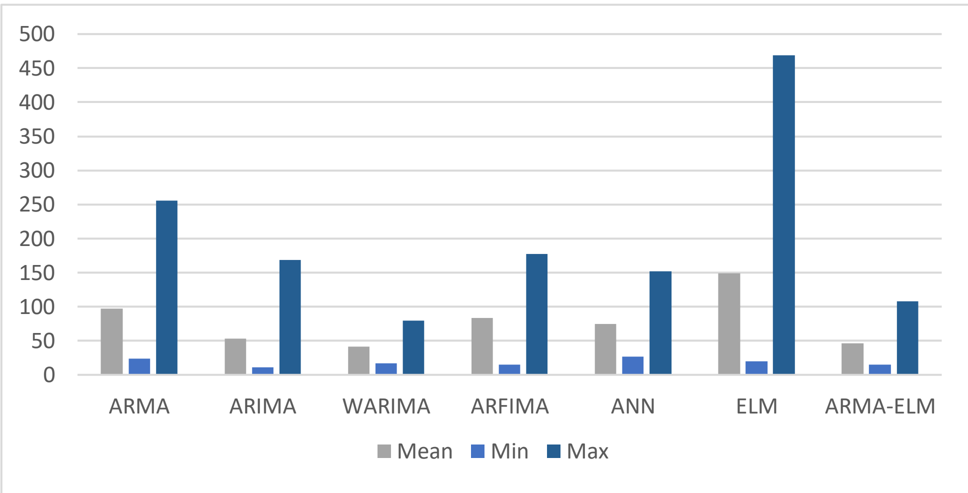
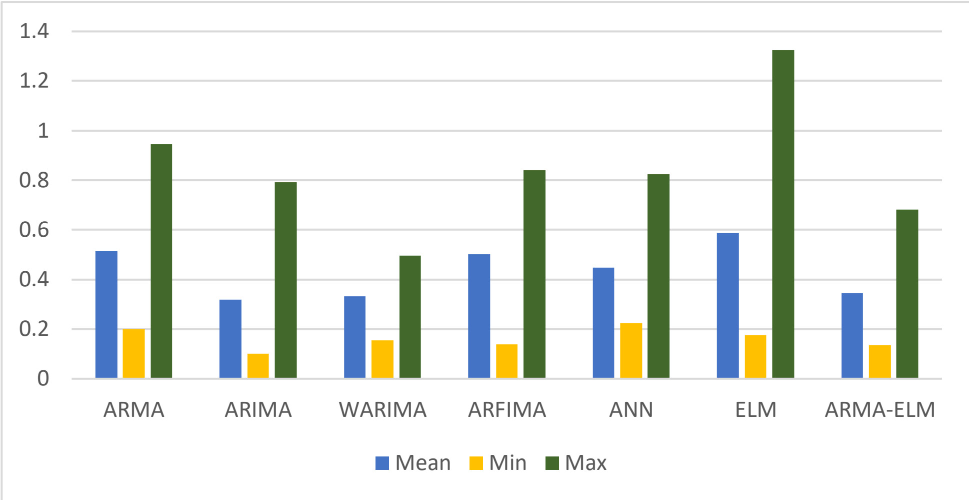


Figure (4.6) SMAPE for Incidences datasets for all countries



Figures (4.7), (4.8), (4.9), and (4.10), respectively display the mean, minimum, and maximum of the RMSE, MAE, MAPE, and SMAPE metrics for the COVID-19 mortalities datasets across all countries.

Figure (4.7) RMSE for Mortalities datasets for all countries

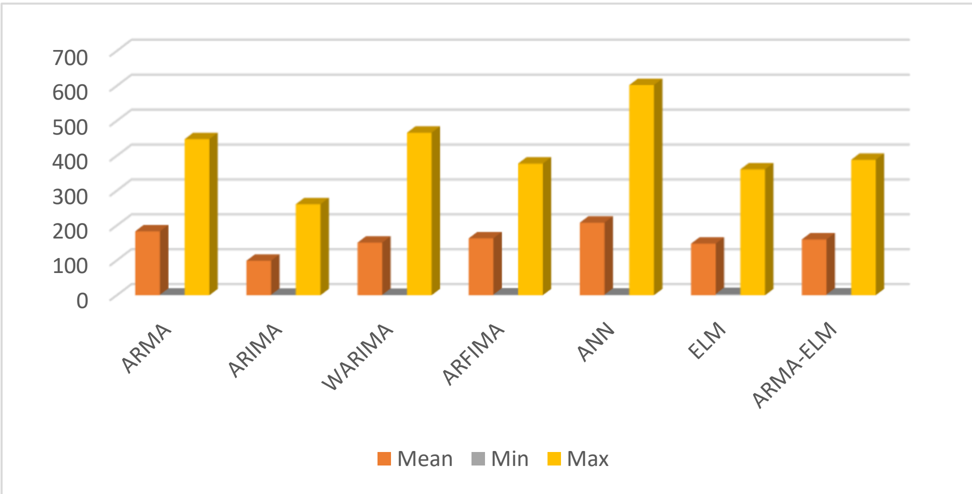


Figure (4.8) MAE for Mortalities datasets for all countries

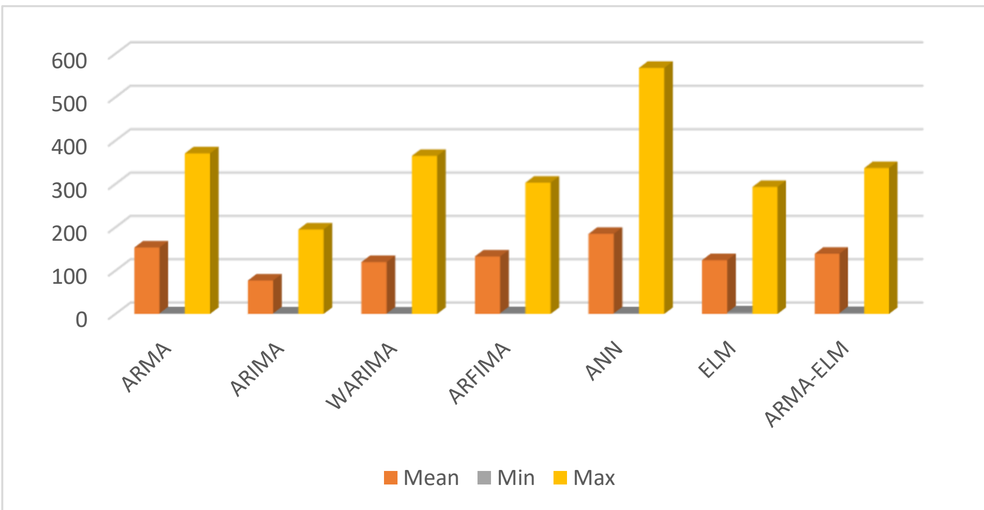


Figure (4.9) MAPE for Mortalities datasets for all countries

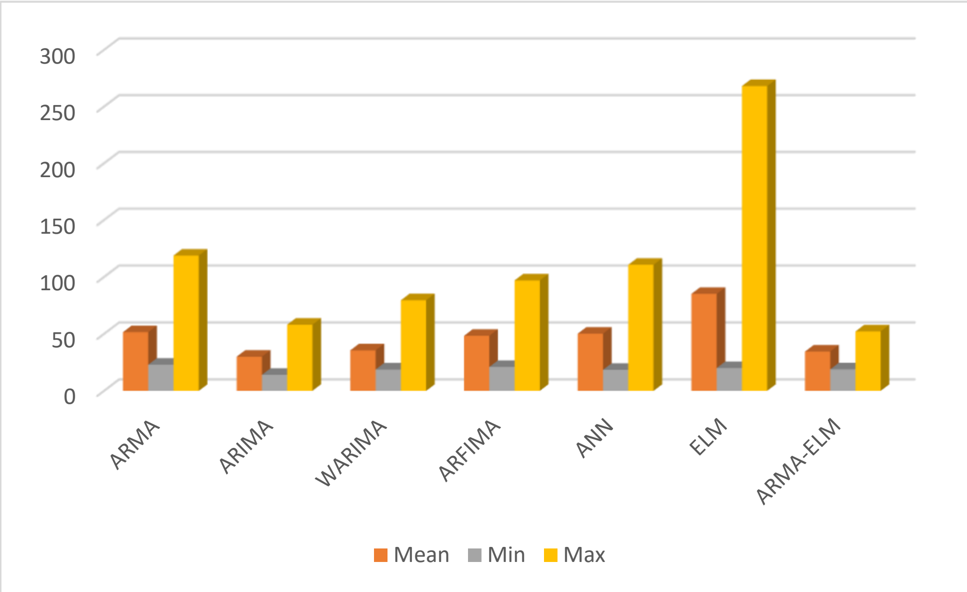
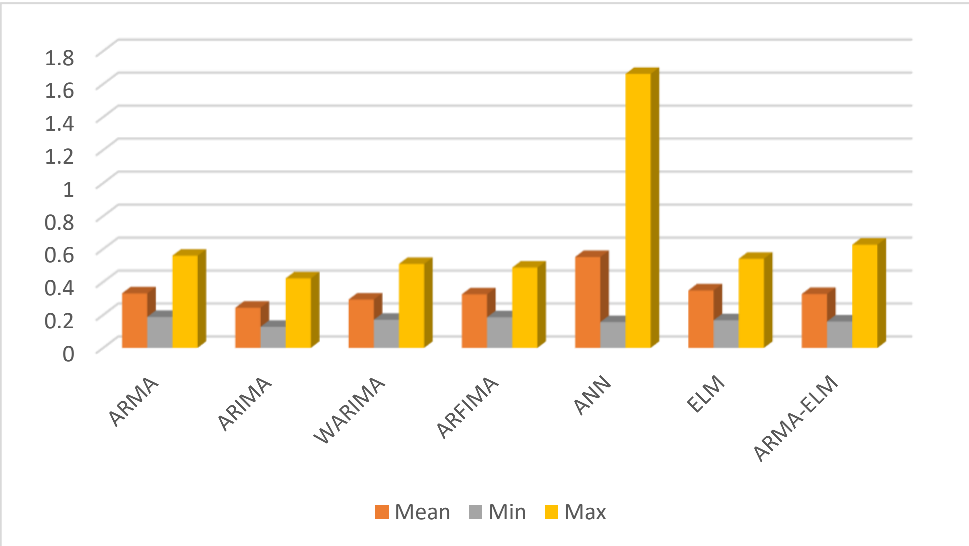


Figure (4.10) SMAPE for Mortalities datasets for all countries



Chapter 5 Discussion

5.1 Discussion

Daily confirmed incidences and mortalities of COVID-19 were used to evaluate multiple statistical time series, shallow neural network, and hybrid combined models across five nations (Brazil, India, Saudi Arabia, Spain, and the USA) form April 2020 to September 2021.Preliminary analysis and inspection of the data indicated substantial differences in both the appearance and correlation structures of these different datasets. In this project seven forecasting models (ARMA, ARIMA, WARIMA, ARFIMA, ANN, ELM, and ARMA-ELM) applied for 7-day ahead projections to the same ten datasets (five incidence datasets, and five mortalities datasets). These models tested using four different model performance metrics (RMSE, MAE, MAPE, and SMAPE). The project results found that there is no specific model has performed the best for all the COVID-19 time series datasets.

Non-linear and non-stationary behaviour were observed to be common features of daily COVID-19 incidences and mortalities records. Table (4.1) shows the COVID-19 time series, ACF, and PACF for daily incidences in five countries (Brazil, India, Saudi Arabia, Spain, and the USA) respectively. The ACF plot for the COVID-19 incidences time series in Brazil, India, Spain, and the USA show several significant lags for example at 7, 14, and 21 lags. These significant might be that the COVID-19 incidences influenced by the action of last three weeks. In contrast, the plot of ACF for the COVID-19 incidences in Saudi Arabia shows that there are only two significant lags at 3, and 5. For the PACF plot of COVID-19 incidences in India, Saudi Arabia, and the USA there is a strong correlation with the adjacent observation where lag = 1. Whereas the PACF plot for the COVID-19 incidences in Brazil has significant lags at 1, 6, and 7. Furthermore, the PACF plot for the COVID-19 incidences in Spain has strong correlation at lag = 1,3,6,8.Table (4.2) shows the COVID-19 time series, ACF, and PACF for daily mortalities in five countries (Brazil, India, Saudi Arabia, Spain, and the USA) respectively. The ACF plot for the COVID-19 mortalities time series in Brazil, and the USA show a few strong significant lags for example at 7, 14, and 21 lags, while in India, and Spain show only one strong correlation at lag = -0.5. Moreover, the ACF plot for the COVID-19 mortalities time series in Saudi Arabia there are several significant lags, but the first two lags are correlated strongly than the rest. For the PACF plot of COVID-19 mortalities in all the five countries has a common strong correlation at lag = 1. Additionally, the PACF plot of COVID-19 mortalities in Brazil, India, Saudi Arabia, and the USA also have significant lags at 6 and 7, 2 and 3, 2, and 5, 6, and 7 respectively. A limitation noted with use of both the ACF and PACF is that the assumption of stationarity is unlikely to hold for all datasets, which could be bias some of the observed ACF and PACF magnitudes.

The models that used in this project responded differently to the COVID-19 time series datasets. For the COVID-19 incidences datasets the ARIMA model performed better in two countries' datasets (Saudi Arabia, and Spain), while the novel ARMA-ELM model has better performance in the other countries (Brazil, India, and the USA) in term of the model performance metrics for 7-days ahead forecasts. On the other hand, for the COVID-19 mortalities datasets the ARIMA model has better performance for the dataset of three counties (Brazil, India, and the USA), whereas the WARIMA model and novel ARMA-ELM model performed better for the data of Saudi Arabia, and Spain respectively based on the model performance metrics for 7-days ahead forecasts. Statistically, the ARIMA model performed better with 50% of the datasets, the hybrid model ARMA-ELM has better performance of 40% of the datasets, and the WARIMA model performed better with only 10% of the datasets. Generally, the forecasting of these datasets show that the statistical stochastic process models perform better than the neural network models. Additionally, the novel ARMA-ELM model also has competitive performance compared to other models such as (ARFIMA, ANN). Another important consideration of these comparisons was revealed in figure (4.3) – (4.6), for the incidence's datasets, the ARMA-ELM had the lowest mean RSME and the maximum RSME as calculated over all 7-days ahead forecasts. A similar result was obtained for the mean MAS and maximum MAS metrics. The WARIMA model performed best for the mean MAPE and maximum MAPE metric but was closely followed by ARMA-ELM (which had a smaller minimum MAPE than WARIMA). However, the ARIMA had the smallest minimum MAPE for all models. For the SMAPE metric, the ARMA-ELM model was outperformed for minimum, mean and maximum statistics but for each case it was nonetheless still within top three performing models. Both RSME and MAE are key metrics of importance when evaluating forecast error. For mortalities datasets, as presented in figures (4.7) – (4.10) the ARIMA model had the lowest mean, minimum, and maximum RMSE. Similarly, the MAE and SMAPE results showed the lowest mean, minimum, and maximum for the ARIMA model. However, the ARIMA model had the smallest mean and minimum MAPE, while the ARMA-ELM model had the lowest maximum value for the same metric. These results suggest that the ARMA-ELM modelling approach is promising for COVID-19 incidence forecasting and might overall the best general performance at an aggregate level (when models are considered globally). Evaluation of model performance for each specific country is still recommended as the ARMA-ELM is not always the top-performing model for each country.

A key finding of this research project is that there are substantial differences in the COVID-19 data correlation structures between different countries. This is potentially due to differences in the underlying transmission dynamics and other medical factors between countries. There was a range of interesting contrasts and comparisons including the correlation lags at lag 7, 14, and 21. These correlated lags might improve the performance of the models specially the ARIMA model.

5.2 Limitations of the Research and Future Directions

This project achieved its objectives through a comprehensive investigation of COVID-19 forecasting model performance across a range of countries and models. In addition, the ARMA-ELM a novel forecasting was evaluated and found to be very promising especially for COVID-19 incidence datasets. Despite the progress achieved in this project , there are limitations which need to be acknowledged. These limitations also indicated future research directions.

- A limited range of statistical and machine learning models were evaluated, further comparisons including other Markov process models and machine learning models such as Support Vector Regression or LSTM (Deep Learning Regression) could have been performed.
- Model assessments and comparisons were still very limited. Only one scenario and a single 7-days ahead forecast was evaluated. More extensive comparisons should include incorporate a running window forecast evaluation, for instance, after a certain minimum amount of data (30 days) models are fit to the data and evaluated for up to 7-days ahead forecasts. The model is then re-trained on 31 days and the forecast errors re-assessed; this process can be repeated throughout. This would be more realistic and comprehensive but would very large amount of computation and currently beyond the scope of the project.
- A major limitation of the ARIMA model is that it assumes a continuous real-valued response, whereas COVID-19 incidence numbers and recorded mortalities are in fact count data.
- Forecast error was evaluated summatively at 7 days. It does not assess the error on each day, some models might be better at shorter range forecast (e.g., 1-day ahead) than others.
- The selection of 7 days ahead forecast was also arbitrary, a one week ahead forecast seems useful but requires further consultation with medical staff and government planners.
- Forecast errors (uncertainty of forecast) were not considered, this could be important in practice. The understanding of acceptable magnitudes of forecast errors is also important in the medical context.

- Data from other counties could also be explored and evaluated to determine if there any particular patterns or characteristics that make certain models more suitable for certain countries.
- The project assumed the recorded incidences and mortalities were perfect records of the situation. This is unlikely in practice and needs to be accounted for in the development and evaluation of novel models.

5.3 Summary

The evaluations of the different forecasting models have been discussed and compared for performance across different countries and for different applications (incidence and mortality forecasting). Both the ACF and PACF analysis suggested the presence of long-range dependencies and that the data correlation structure differed between countries. A substantial limitation of both the ACF and PACF is the assumption of data stationarity, which appears unlikely to hold for these datasets. Nonetheless, there is some evidence to indicate long-range dependencies in these datasets. Utilising such long-range dependencies was not deemed practical due to the long duration required before sufficient data could be collected in order to estimate the model parameters.

A total of 7 forecasting models were evaluated on COVID-19 datasets from 5 countries. Model performance varied dependant on country and application (incidence or mortality) forecasting. In all scenarios examined, it is recommended to evaluate a range of models to determine the best performing model for each country and application. In terms of general recommendations, both the ARIMA and the ARMA-ELM models were top-performing models, the ARMA-ELM was most promising for the COVID-19 incidence datasets, whilst the ARIMA was best for the COVID-19 mortality datasets. Both of these models should be first choices if a comprehensive comparison of model performance is not possible when developing country specific COVID-19 forecasts.

Chapter 6 Conclusion

6.1 Conclusion

Statistical stochastic process time series models and neural network models have played an important pivotal role in short-term forecasting of pandemics. The project has focussed on pre-existing models (ARMA, ARIMA, WARIMA, ARFIMA, ANN and ELM). In addition to these models the project has explored a novel model which combined between a statistical stochastic process time series model with a neural network to produce the ARMA-ELM. Seven different models used to forecast the COVID-19 daily confirmed incidence and mortalities datasets for five countries (Brazil, India, Saudi Arabia, Spain, and USA) from April 2020 to September 2021. A 7-days ahead forecast for these time series datasets used to evaluate and compare the model performance metrics (RMSE, MAE, MAPE, and SMAPE) for all counties datasets. The project results found that for the COVID-19 incidences datasets, the ARIMA and ARMA-ELM models overall have better model performance metrics than other models such as WARIMA and ANN. The ARMA-ELM was found to be very competitive when evaluating model performance metrics at a global scale instead of at a country specific level. The COVID-19 mortalities forecast indicated that ARIMA model performed the best overall (smallest error metrics for Brazil, India, and the USA). In contrast, according to the model performance metrics, the COVID-19 mortalities forecasts in Saudi Arabia and Spain were best described using the WARIMA and ARMA-ELM models respectively. Overall, none of the forecasting models performs equally well across all datasets, highlighting the need to explore a range of models when developing country-specific forecasts. Future research should validate the reliability of the COVID-19 datasets and address any uncertainties in these data sets. More extensive investigations and assessments beyond this scenario presented should also be conducted to develop a more comprehensive understanding of model performance.

6.2 Reference Lists

Alzahrani, SI, Aljamaan, IA & Al-Fakih, EA 2020, ‘Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions’, *Journal of Infection and Public Health*, vol. 13, pp. 914-919.

Armachie, J 2017, ‘Arbitrage Opportunity In The Ghanaian Stock Market An ARFIMA Approach’, PhD thesis, The University of Ghana, Legon Boundary, Accra, Ghana.

Box, GEP, Jenkins, GM, Reinsel, GC & Ljung, GM 2015, *Time series analysis: forecasting and control,* 5th edn, John Wiley & Sons.

Brunton, SL & Kutz, JN 2022, *Data-driven science and engineering: Machine learning, dynamical systems, and control,* 2nd edn, Cambridge University Press.

Chai, T & Draxler, RR 2014, ‘ Root mean square error (RMSE) or mean absolute (MAE)? – Arguments against avoiding RMSE in the literature’, *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247- 1250.

Chakraborty, T & Ghosh, I 2020, ‘Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis’, *Chaos, Solitons and Fractals*, vol. 135, p. 10.

Chakraborty, T, Ghosh, I, Mahajan, T & Arora, T 2022, ‘Nowcasting of COVID-19 confirmed cases: Foundations, trends, and challenges’, *Modeling, Control and Drug Development for COVID-19 Outbreak Prevention,* pp. 1023-1064.

Chen, JX 2019, ‘Applications of Extreme Learning Machines’, *Computing in Science & Engineering,* pp. 4–5.

Chintalapudi, N, Battineni, G & Amenta, F 2020, ‘COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: A data driven model approach’, *Journal of Microbiology, Immunology and Infection*, vol. 53, no. 3, pp. 396-403.

Contreras, J, Espinola, R, Nogales, FJ & Conejo, AJ 2003, ‘ARIMA models to predict next-day electricity prices’, *IEEE Transactions on Power Systems*, vol. 18, no. 3, pp. 1014-1020.

Cucinotta D, Vanelli, M 2020, *WHO Declares COVID-19 a Pandemic*, viewed 11 October 2020, <https://pubmed.ncbi.nlm.nih.gov/32191675/>.

Davidescu, AA, Apostu, SA & Marin, A 2021, ‘Forecasting the Romanian Unemployment Rate in Time of Health Crisis—A Univariate vs. Multivariate Time Series Approach’, *International Journal of Environmental Research and Public Health,* vol. 18, no. 21, p. 31.

Dürre, A, Fried, R & Liboschik, T 2015, ‘Robust estimation of (partial) autocorrelation’, *Wiley Interdisciplinary Reviews: Computational Statistics,* vol. 7, no. 3, pp. 205-222.

Faraway, J & Chatfield, C 1998, 'Time Series Forecasting with Neural Networks: A Comparative Study Using the Airline Data', *Journal of the Royal Statistical Society,* vol. 47, no. 2, pp. 231-250.

Fattah, J, Ezzine, L, Aman, Z, Moussami, HE & Lachhab, A 2018, 'Forecasting of demand using ARIMA model'*, International Journal of Engineering Business Management,* vol. 10, p. 9.

Flores, JHF, Engel, PM & Pinto, RC 2012, 'Autocorrelation and partial autocorrelation functions to improve neural networks models on univariate time series forecasting', in The 2012 *International Joint Conference on Neural Networks (IJCNN),* Brisbane, Australia, p. 8.

Hilal, YY, Yahya, A, Ishak, WW & Asha'ari, ZH 2020, 'Genetic Algorithms and Artificial Neural Networks for Optimization Palm Oil Production Models Based on Various Parameter', *The Scientific Journal of King Faisal University*, p. 13.

Huang, GB, Zhu, QY & Siew, CK 2006, 'Extreme learning machine: Theory and applications', *Neurocomputing,* vol. 70, no. 1, pp. 489-501.

Huang, J, Yu, ZL & Gu, Z 2018, 'A clustering method based on extreme learning machine', *Neurocomputing*, vol. 277, pp. 108-119.

Hyndman, RJ & Athanasopoulos, G 2018, *Forecasting: principles and practice*, 2nd edn, Monash University, Melbourne, Australia.

Hyndman, RJ & Khandakar, Y 2008, 'Automatic Time Series Forecasting: The forecast Package for R', *Journal of Statistical Software,* vol. 27, no. 3, p. 22.

Jourdain, B & Lelievre, T 2003, 'Mathematical analysis of a stochastic differential equation arising in the micro-macro modelling of polymeric fluids', *In Probabilistic methods in fluids,* pp. 205-223.

Jurado, S, Mugica, F, Nebot, A & Avellana, N 2013, 'Towards the Development of the Smart Grid: Fast Electricity Load Forecasting Using Different Hybrid Approaches', *Artificial Intelligence Research and Development, IOS Press,* 2013, pp. 185-188.

Khan, F, Saeed, A & Ali, S 2020, 'Modelling and forecasting of new cases, deaths and recover cases of COVID-19 by using Vector Autoregressive model in Pakistan', *Chaos, Solitons & Fractals,* vol. 140, p. 5.

Khan, MA, Majid, A, Akram, T, Hussain, N, Nam, Y, Kadry, S, Wang, SH & Alhaisoni, M 2021, 'Classification of COVID-19 CT Scans via Extreme Learning Machine', *Computers, Materials & Continua,* vol. 68, no. 1, pp. 1003-1019.

Kourentzes, N 2022, 'Package 'nnfor'', p. 21.

Lejeune, E & Linder, C 2020, 'Interpreting stochastic agent-based models of cell death', *Computer Methods in Applied Mechanics and Engineering,* vol. 360, p. 21.

Liu, K, Chen, YQ & Zhang, X 2017, 'An evaluation of ARFIMA (Autoregressive Fractional Integral Moving Average) Programs, *Axioms*, vol. 6, no. 2, p. 16.

Li, K & Huang, R 2022, 'A CKKS-based Privacy Preserving Extreme Learning Machine', *International Journal of Network Security*, vol. 24, no. 1, pp. 166-175.

Maier, HR & Dandy, GC 2000, 'Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications', *Environmental Modelling & Software,* vol. 15, no. 1, pp. 101-124.

Madhav, N, Oppenheim, B, Gallivan, M, Mulembakani, P, Rubin, E & Wolfe, N 2018, 'Pandemics: risks, impacts, and mitigation', in DT, Jamison, H, Gelband, S, Horton, P, Jha, R, Laxminarayan, CN, Mock & R, Nugent (eds), Disease Control Priorities: Improving Heath and Reducing Poverty, World Bank Group, Washington, USA, pp. 315-345.

Masa, AS & Diaz, JFT 2017, 'Long-memory Modelling and Forecasting of the Returns and Volatility of Exchange-traded Notes (ETNs)', *Margin-The Journal of Applied Economic Research,* vol. 11 no. 1, pp. 23-53.

Mélard, G & Pasteels, JM 2000, 'Automatic ARIMA modeling including interventions, using time series expert software', *International Journal of Forecasting*, vol. 16, no. 4, pp. 497-508.

Niazkar, HR & Niazkar, M 2020, 'Application of artificial neural networks to predict the COVID-19 outbreak', *Global Health Research and Policy,* vol. 5, no. 1, p. 11.

Paul, RK Samanta, S 2017, 'Package 'WaveletArima'', p. 5.

Pavlicko, M, Vojteková, M & Blažeková, O 2022, 'Forecasting of Electrical Energy Consumption in Slovakia', *MDPI,* vol. 10, no. 4, p. 20.

Qing, Y, Zeng, Y, Li, Y & Huang, GB 2020, 'Deep and wide feature based extreme learning machine for image classification'*, Neurocomputing,* vol. 412, pp. 426-436.

Rajpal, S, Agarwal, M, Rajpal, A, Lakhyani, N, Saggra, A & Kumar N 2022, 'COV-ELM classifier: An extreme learning machine based identification of COVID-19 using chest X-ray images', *Intelligent Decision Technologies,* vol. 16, pp. 193-203.

Ribeiro, MHDM, Silva, RGD, Mariani, VC & Coelho, LDS 2020, 'Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil', *Chaos, Solitons and Fractals,* vol. 135, p. 10.

Shi, Y, Chen, G & Li, J 2018, 'Malicious Domain Name Detection Based on Extreme Machine Learning', *Neural Processing Letters,* vol. 48, pp. 1347-1357.

Teräsvirta, T, Dijk, DV & Medeiros, MC 2005, 'Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic tome series: A re-examination', *International Journal of Forecasting,* vol. 21 no. 4, pp. 755-774.

Tong, H 1990, Non-linear time series: a dynamical system approach, Oxford University Press.

Wang, J, Lu, S, Wang, SH & Z, YD 2021, 'A review on extreme learning machine', *Multimedia Tools and Applications,* 22 May, p. 50.

Wang, W, Tang, J & Wie, F 2020, ' Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China', *Journal of Medical Virology Wiley*, vol. 92, no. 4, pp. 441-447.

World Health Organization 2020, COVID-19 – China, Viewed 14 October 2020, <https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON229>.

World Health Organization 2020, WHO statement on novel coronavirus in Thailand, Viewed 11 October 2022, <https://www.who.int/news/item/13-01-2020-who-statement-on-novel-coronavirus-in-thailand>.

World Health Organization 2020, WHO Coronavirus (COVID-19) Dashboard, Viewed 20 February 2022, < https://covid19.who.int>.

Yang, Q, Wang, J, Ma, H & Wang, X 2020, 'Research on COVID-19 based on ARIMA model-Taking Hubei, China as an example to see the epidemic in Italy', *Journal of Infection and Public Health,* vol. 13, pp. 1415-1418.

Yang, Y, Peng, F, Wang, R, Guan, K, Jiang, T., Xu, G, Sun, J & Chang, C 2020, 'The deadly coronaviruses: The 2003 SARS pandemic and the 2020 novel coronavirus epidemic in China', *Journal of Autoimmunity,* vol. 109, p. 16.

Zięba, A & Ramza, P 2011, 'Standard deviation of the mean of autocorrelated observations estimated with the use of the autocorrelation function estimated from the data', *Metrology and Measurement Systems*, vol. 18 no. 4, pp. 529-542.

Zhang, L, Liu, LJ, Bao, SN, Qiang, MT & Zou XM 2015, 'Transportation Mode Detection Based on Permutation Entropy and Extreme Learning Machine', *Mathematical Problems in Engineering,* vol. 2015, p. 10.