**Big Data Analytics Predictive Modeling of COVID-19 Cases and Deaths in the United States.**

**NAME: OSEMEIGAH IGBAFE**

**STUDENT NUMBER: 2046969**

**The University of Wolverhampton**

1. **ABSTRACT**

The COVID-19 outbreak has resulted in a worldwide public health calamity, affecting millions of people. For guiding decision-making and policy creation, accurate surveillance and analysis of the virus's circulation are crucial. In many regions of the world, the pandemic has created an urgent need for reliable and up-to-date information on confirmed cases, fatalities, and testing rates. The goal of this study is to evaluate and analyse the COVID-19 dataset to get insight into the virus's spread and impact on various nations and places. The study investigates and delivers insights from COVID-19 data using big data analytics approaches. To predict the spread and impact of the infection, the approach utilises data pre-processing, exploratory data analysis, and machine learning techniques. The findings of the report can help governments and healthcare professionals devise appropriate COVID-19 pandemic response strategies. It can also help researchers gain a better understanding of the virus's dissemination and impact, which can lead to greater research in this area.

1. **INTRODUCTION**

The COVID-19 outbreak has affected millions of people worldwide, putting a significant strain on many nations' healthcare systems. In this circumstance, the need for a trustworthy data analysis system becomes important. The collection contains statistics on COVID-19 cases reported in many states around the United States, including the number of confirmed cases, deaths, testing rates, and incidence rates. Big Data technology's

**NAME: Amir, Shahzeb**

**The University of Wolverhampton**

implementation in healthcare has resulted in substantial break throughs in data analysis and management. Sentiment analysis, NoSQL databases, data warehousing, and data mining are all important Big Data technology components employed in the healthcare industry. These technologies can aid in the discovery of patterns and trends in data, offering useful insights for healthcare decision-making. Previous studies relied solely on historical and observational data. To create predictions, these studies employ statistical and mathematical models. The sample size in this research is small. These studies only consider one or two aspects of data. The globe needs a more efficient and accurate forecast system capable of analysing historical and real-time data from around the world. The new technique should be more adaptive and user-friendly in terms of importing all data connected to the research topic.

In this work, we use Python programming to undertake a comprehensive study of the COVID-19 dataset. The study covers a wide range of data analysis topics, such as data cleansing, exploration, visualisation, and modelling. The research also investigates the use of various Big Data technologies in healthcare and their potential influence on improving healthcare outcomes.

The following contributions have been made by this work:

* A brief summary of big data technology and its many components
* Examining uses of big data technologies such as data warehousing, data mining, sentiment analysis, and NoSQL.
* A case study on COVID-19 data is offered to demonstrate the practical applications of big data technologies.
* In order to improve COVID-19 prediction accuracy in the United States.
* To assess the efficacy of several machine learning models in COVID-19 forecasting.
* To give information on the function of non-pharmaceutical interventions (NPIs) in COVID-19 spread.
* Using big data to assist the government and healthcare practitioners in making informed decisions by giving fast and accurate estimates.

1. **Related Work**

Several study suggestions in the active literature concentrate on massive COVID-19 data analytics, the scientific subject closest to our research. Here are some of the most notable.

[1] Investigates the problem of predicting the progression of Covid-19 transmission using deep hybrid learning on massive volumes of social media data. A deep hybrid learning model known as (ODANN) optimised data assimilated neural network built on (NN) neural networks coupled with assimilation of data and natural language processing (NLP) extraction feature methods used to simulate the pandemic, daily concurrent COVID-19 time-series recordings, as well as massive amounts of COVID-19 related Twitter data, were processed.

[2] focuses on modelling and tracking COVID-19 cases using significant data analytics approaches built on top of high-performance computing cluster (HPCC) technology. The authors demonstrate how they created a Corona spread model using unique, extensive data analytics methodologies and technology. They applied their understanding of Ebola transmission models to efficiently duplicate Corona spread, revealing new insights and contributing to reducing Corona cases.

[3] delves deeply into the COVID-19 categorisation issue. The authors note that many researchers have done very well in developing ground-breaking deep learning (DL) models for the automated screening of COVID-19 using computerised tomography (CT) images. However, concerns remain concerning how tiny perturbations and structural changes in CT scans affect performance stability. The study proposes a feature extraction strategy that combines a moment invariant (MI) method with a deep learning (DL) algorithm to overcome instabilities in existing COVID-19 classification models. A cascade fusion approach was used; the suggested method combines MI-based features into the DL models. It was discovered that combining MI and DL characteristics has the potential to increase the sensitivity and accuracy of the COVID-19 classification.

Researchers have worked hard [4], [5] to investigate the influence of extrinsic triggers on COVID-19 mortality rates and case spread, such as weather [6] and air pollution [5], [7], [8]. [9] We explored how gender, birth year, infection grounds, and geography affected the reported number of recovered and deceased COVID-19 patients.

Diagram

Description automatically generated

Fig. 1 SPARK work on big data

Our research, on the other hand, looks at the relationship between COVID-19 cases, deaths, and hospitalisations in various US states over time. We employ a big dataset, as well as numerous data analytic and visualisation tools, to offer a complete picture of how these traits evolved during the epidemic. Furthermore, we investigate the effects of various interventions, such as lockdowns and vaccines, on the patterns of these variables. This one-of-a-kind technique generates useful information that may be utilised to affect public health policies and practises.

1. ***Methodology***

Timely action and confidence in scientific truths are necessary to limit the impact of climate change on COVID-19 death rates, and case spread [10]. Fig. 1 depicts a high-level overview of the system provided in this paper for colossal data processing using Spark. This image is divided into three primary phases: data collection, spatial and temporal analysis with Spark, and visualisation, the most crucial stage in exhibiting consistent findings.

The number of confirmed cases, incident rate, and fatalities are the three primary factors in this study. This research primarily focuses on constructing statistical regression models for anticipating the outbreak's incidence, peak, and trend.

4a**. Proposed Model for Autoregressive Integrated Moving Average Forecasting**

In real-time, the ARIMA model may predict and evaluate various non-stationary time series problems, including socioeconomic, commercial, and epidemiological studies [11]. For example, the non-seasonal ARIMA model is used in this study to analyse the incidence of the COVID-19 pandemic in twenty-one badly impacted United States from March 18th, 2020, to March 31st, 2021, as well as the top six worst-affected nations, including India, from January 30th, 2020, to March 31st, 2021. ARIMA (p, d, q) is the name given to this model because it has three components: an autoregressive component (AR), a moving average contribution (MA), and an integration component (I).

[10] A linear combination of the previous p observations and a random error with a fixed value yields the mathematical formula for the model's autoregressive term (AR) with a lag of p. This may be written as:

(1)

Where yt and ϵt are the target value and random error at period t, and φi(i=1,2,...,p) are the model parameters with a constant c. The integer constant p represents the model's order.

MA(q), the model's moving average component, uses prior errors as explanatory variables and is formally defined as:

(2)

Where is the series mean, (j=1,2,...,q) is a collection of values representing the model parameters, while q denotes the model's order. The random error (i.i.d.) is a sequence of independent and identically distributed random variables with a mean of zero and a constant variance.

These two models are merged to generate the ARIMA model, which has the following mathematical formulation [12]:

(3)

The following parameters are defined here:

* is a constant term, p0 is the order of the AR model, and AR(p) is the number of lags. The difference in degree is denoted by d0, while the integration parameter is denoted by I(d).
* MA(q) denotes the number of lags, whereas q0 is the order of the MA model.
* The AR(p) model parameters are ,(i=1,2,...,p).
* MA(q) model parameters are {θj},(j=1,2,...,q).
* The random error is represented by

The differencing parameter is critical for increasing the series' stationarity and, as a result, dropping the mean to zero. It may be stated mathematically as follows:

(4)

The parameter estimation, model development, and forecasting of time series datasets require the four computational operations listed below:

* **Transformation stage:** If the presentation of a time series dataset implies that it is non-stationary, the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) diagnostic tool [11] [15] can be used to confirm its consistency. The finite difference transformation approach, which is mentioned in equation (4) [12], is then used to conduct the appropriate transformations on the series stated in equation (3) to generate the time series while assuring that it is stationary. Following the change, the series is re-evaluated using KPSS to determine whether it is stationary around the mean.
* **Identification Model:** A stationary series is obtained in the initial stage of the modelling procedure. The purpose of the second step is to find the optimal ARIMA (p, d, q) model for by analysing the autocorrelation function (ACF) and partial autocorrelation function (PACF) [12], [13]. This is an important stage in the model's development since it includes selecting many viable model structures for future investigation. As a result, further study is frequently required at this phase to narrow down the probable selection of the best model from among the candidate models in the series.
* **Estimate Model:** In this phase, all ARIMA models examined in the previous stage are estimated to select the best model for the series. Then, the Box-Ljung test [14] may be used to pick the model that best fits the series, and the model parameters can be calculated using a conditional sum of square likelihood. This test determines if the mistakes are random or independently and identically distributed (i.i.d).
* **Diagnostic Check:** The autocorrelation function (ACF) and partial autocorrelation function (PACF) of the residuals are analysed during the diagnostic check step to validate that the autocorrelations of the error are very low and that the model is a good fit for the series. The predictability of the model is then assessed by computing the mean absolute percentage error (MAPE), mean squared error (MSE), and root mean squared error (RMSE).

The model's workflow diagram, shown in Fig 2, reflects the four previously stated procedures. Following bias and variance error reduction, 80% of the ill data is utilised for the trained model, with the remaining 20% used to verify and forecast the pandemic's future values. The model anticipates 90-day future values for time-series datasets.

Diagram

Description automatically generated

Fig

1. **Result and Discussion**
2. ***Dataset***

A dataset is a grouping of data that is often maintained and organised in the form of a database or spreadsheet. A dataset is a collection of data points or records, each with information about a specific item or observation. Datasets are widely used to analyse, analyse, and predict data in a variety of disciplines, including scientific research, industry, and government. They might range from a few records to millions or billions.

This part of the report acknowledges the dataset's owner that grants access to researchers.

1. ***Dataset Details***

The data includes COVID-19 incidents, deaths, and hospitalisations from January 21, 2020, to January 2, 2022, in several US states. The data was gathered from the COVID-19 Data Repository by the Johns Hopkins University Centre for Systems Science and Engineering (CSSE). It publishes daily data on confirmed cases, fatalities, and hospitalisations, as well as aggregate statistics on cases, fatalities, and hospitalisations. The data also includes each state's latitude and longitude, which may be used for geographical analysis. The dataset is in CSV format and has 729 rows and 45 columns.

Table

Description automatically generated with medium confidence

Fig 3

Text, table

Description automatically generated

Fig 3b

Fig 3 depicts a dataset with 58 Pandas data frames and 21 columns. The columns contain data such as province/state, country/region, latitude and longitude, confirmed cases, fatalities, FIPS code, incident rate, total test results, case fatality ratio, UID (Unique Identifier), ISO3 code, testing rate, and date. Some columns, such as recovered, active, people hospitalised, hospitalisation rate, individuals tested, and fatality rate, have no non-null values. Float64, int64, and object are examples of column data types. While in figure 2d it shows the description of the Covid-19 dataset.

1. ***Cleaning Data***

This process includes data cleansing and modification to prepare the dataset for analysis. The data had missing numbers, anomalies, and inconsistencies. By converting all category variables into numerical variables, the data was also changed into an analysis-ready format.

Text

Description automatically generated

Fig 4. Cleaned dataset.

1. ***Data Analysis***

At this stage, the dataset was analysed using a variety of statistical and machine learning methodologies. This included looking at the relationships between numerous components, identifying patterns and trends, and producing insights into the disease's spread.

1. ***Visualisation of Data***

The data analysis results are visualised using various graphs and charts to provide a clear image of the findings. is the use of visual components such as graphs, charts, and maps to display information gained from data mining, assisting in the interpretation of enormous volumes of data and enabling decision-making based on it. Machine learning simplifies difficult research tasks such as predictive modelling, which may subsequently be used to generate successful communication visualisations.

Diagram

Description automatically generated

*Figure 4 analysis and visualisation of data*

Chart

Description automatically generated

1. ***Experimental Method***

Our significant research emphasizes the computational approach for constructing the optimal ARIMA (p, d, q) model for time series data gathered in the United States. Therefore, only the optimal order model that satisfies the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) value, p-value from the Box-Ljung test on residuals, and predicted root-mean-square error (RMSE) are provided for other states within the US and countries outside the US. The best models are then used to forecast the occurrence of disease in the United States and five other nations worldwide. Furthermore, the model for the six most affected states predicts illness transmission.Chart

Description automatically generated

***CONCLUSION***

Finally, our examination of the association between COVID-19 incidences, fatalities, and hospitalisations in various US states, as well as how they have escalated over time, revealed a substantial link between the number of verified cases and the number of deaths. We also discovered that death rates varied greatly between states, with some having considerably higher mortality rates than others. Furthermore, as the number of cases grows, so does the number of hospitalisations, putting a burden on healthcare systems.

We used publicly available data from Johns Hopkins University's Centre for Systems Science and Engineering (CSSE) to perform our research. This information included daily updates on COVID-19 cases and fatalities across the United States. We used a variety of strategies to analyse this data, including data cleansing, exploratory data analysis, and regression analysis.

Our findings have important implications for governments and healthcare practitioners dealing with the COVID-19 outbreak. Understanding the association between cases, fatalities, and hospitalisations helps governments to better allocate resources to regions most impacted by the virus, while healthcare personnel can plan for surges in hospitalisations and assure proper treatment.

Overall, our research adds to the expanding body of knowledge about the COVID-19 pandemic and emphasises the need of employing data-driven techniques to understand and respond to the epidemic.

1. ***Reference***
2. Alvin Wei Ze Chew, Yue Pan, Ying Wang and Limao Zhang, "Hybrid Deep Learning of Social Media Big Data for Predicting the Evolution of COVID-19 Transmission", *Knowl. Based Syst.*, vol. 233, pp. 107417, 2021.
3. Flavio Villanustre, Arjuna Chala, Roger Dev, Lili Xu, Jesse Shaw, Borko Furht, et al., "Modeling and Tracking COVID-19 Cases Using Big Data Analytics on HPCC System Platform", *J. Big Data*, vol. 8, no. 1, pp. 33, 2021.
4. Ervin Gubin Moung, Chong Joon Hou, Maisarah Mohd Sufian, Mohd. Hanafi Ahmad Hijazi, Jamal Ahmad Dargham and Sigeru Omatu, "Fusion of Moment Invariant Method and Deep Learning Algorithm for COVID-19 Classification", *Big Data Cogn. Comput.*, vol. 5, no. 4, 2021.
5. A. L. R. Ribeiro and N. W. A. Sousa, "Besides the climate model other variables driving the covid-19 spread in brazil", *The Science of the Total Environment*, vol. 737, pp. 140211, 2020
6. J. D. Berman and K. Ebisu, "Changes in us air pollution during the covid-19 pandemic", *Science of the Total Environment*, vol. 739, pp. 139864, 2020.
7. A. Auler, F. Cássaro, V. da Silva and L. Pires, "Evidence that high temperatures and intermediate relative humidity might favor the spread of covid-19 in tropical climate: A case study for the most affected brazilian cities", *Science of The Total Environment*, pp. 139090, 2020.
8. K. Chen, M. Wang, C. Huang, P. L. Kinney and P. T. Anastas, "Air pollution reduction and mortality benefit during the covid-19 outbreak in china", *The Lancet Planetary Health*, vol. 4, no. 6, pp. e210-e212, 2020.
9. F. Dutheil, J. S. Baker and V. Navel, "Covid-19 as a factor influencing air pollution?", *Environmental Pollution (Barking Essex: 1987)*, vol. 263, pp. 114466, 2020.
10. N. Al-Rousan and H. Al-Najjar, "Data analysis of coronavirus covid-19 epidemic in south korea based on recovered and death cases", *Journal of Medical Virology*, 2020.
11. R. D. Manzanedo and P. Manning, "Covid-19: Lessons for the climate change emergency", *Science of the Total Environment*, vol. 742, pp. 140563, 2020.
12. R. Gupta and S. K. Pal, "Trend analysis and forecasting of COVID-19 outbreak in India", *medRxiv*, Mar. 2020.
13. D. Ray, M. Salvatore, R. Bhattacharyya, L. Wang, J. Du, S. Mohammed, et al., "Predictions role of interventions and effects of a historic national lockdown in India’s response to the COVID-19 pandemic: Data science call to arms", *Harvard Data Sci. Rev.*, vol. 2020, 2020.
14. M. Bhatnagar, "COVID-19: Mathematical modeling and predictions", Apr. 2020.
15. T. Rahman, A. Akinbi, M. E. H. Chowdhury, T. A. Rashid, A. Şengür, A. Khandakar, et al., "COV-ECGNET: COVID-19 detection using ECG trace images with deep convolutional neural network", *arXiv:2106.00436*, 2021, [online] Available: <https://arxiv.org/abs/2106.00436>.
16. F. S. Ahmad, L. Ali, H. A. Khattak, T. Hameed, I. Wajahat, S. Kadry, et al., "A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (EHRs)", *J. Ambient Intell. Hum. Comput.*, vol. 12, no. 3, pp. 3283-3293, Mar. 2021