International Journal of Computer Science and Engineering Research and Development (IJCSERD)
ISSN Print: 2248-9363, **ISSN Online:** 2248-9371
Volume 6 Issue 1, January–April (2023), pp. 1-8

SURVEY ON PANDEMIC OUTBREAK PREDICTION METHODS

M. Rohit
Department of Computer Science and Engineering,
SRM University, Kattankulathur – 603203, Tamil Nadu, India

Dr. E. Poovammal
Department of Computer Science and Engineering,
SRM University, Kattankulathur – 603203, Tamil Nadu, India

ABSTRACT

The emergence of COVID-19 in 2019 led to a global pandemic, causing significant damage to society, politics, and the economy, with millions of lives lost. Accurately forecasting COVID-19's future spread is critical in managing its impact. This study evaluates the performance of four time-series analysis models - ARIMA, Prophet, LSTM, and Transformer models - to predict future COVID-19 trends in six nations. We obtain COVID-19 case data from the publicly available database of Johns Hopkins University Center for Systems Science and Engineering and use it to make predictions repeatedly across all models. Performance evaluation is done using mean squared error (MSE) and mean absolute error (MAE). Our findings show that the LSTM model achieves the lowest MSE and MAE across all countries. Although the Transformer model performs poorly overall, it has the second-best performances in certain countries. These results highlight the high accuracy of the LSTM model in forecasting the spread of COVID-19, enabling countries to better plan and implement measures to control the virus.

Keywords: *Temporal Pattern Recognition, Autoregressive Integrated Moving Average, Support Vector Machine, and Transform Transformer.*

Cite this Article: M. Rohit and Dr. E. Poovammal, Survey on Pandemic Outbreak Prediction Methods, International Journal of Computer Science and Engineering Research and Development (IJCSERD), 6(1), 2023, pp. 1-8.

1. INTRODUCTION

This chapter serves as an introduction to the thesis project, providing an overview for the reader. The content is organized into two sections. The first section focuses on the coronavirus disease (COVID-19) and examines the latest methods for analyzing and predicting time series. In the second section, we present the research problem that will be investigated in this study.

The study's aims and methodology are then discussed in depth. The chapter concludes with a presentation of the thesis's organisational structure and a brief discussion of its delimitations.

1.1. BACKGROUND

In December 2019, a new outbreak of a respiratory illness surfaced in Wuhan, China. This illness was later identified as COVID-19, caused by the SARS-COV-2 virus. Patients exhibited symptoms of severe pneumonia, which can lead to organ failure and death. The virus is highly contagious and primarily transmitted through respiratory droplets, making it easily transmissible if necessary precautions are not taken. Due to its highly contagious nature, COVID-19 quickly became a global pandemic, causing over 214 million confirmed cases and claiming the lives of over four million individuals worldwide [1]. The pandemic has had a profound impact on individuals' physical and emotional health, daily routines, and global economies. Governments worldwide have been grappling with the challenges of implementing measures such as mandating mask-wearing, travel restrictions, and school closures to contain the spread of the disease and "flatten the curve" [2].

1.2. CONFORMITY WITH THE CURRENT STANDARD

Various methods exist for analysing time-series data, such as the total number of COVID-19 cases or any other disease, to identify patterns and predict future trends. These methods encompass statistical techniques like the Autoregressive Integrated Moving Average (ARIMA) model, as well as Machine Learning (ML) and Deep Learning (DL) techniques, including the Long Short-Term Memory (LSTM).

1.3. PROBLEM

Although time-series analysis methods have shown promising results in the study of infectious disease, it remains difficult to find an appropriate way to analyse COVID-19, which is one of the most serious, deadliest health crises in recent years and has more impact in every aspect of society worldwide.

To address the stated problem, certain scientific issues need to be addressed. The first issue of paramount importance is the interpretation of case numbers. A range of statistics, such as current cases, cumulative cases, daily reports, recovered patients, and death toll, are available. It is crucial to determine which statistics are crucial in predicting the trend of COVID-19 while maintaining high accuracy.

1.4. PURPOSE

As this is now a major global concern, it is essential to gain a thorough understanding of the COVID-19 curve and predict its future trend. Doing so will allow us to better understand the disease itself and to obtain an approximate idea of the current severity of the pandemic, allowing for the implementation of appropriate measures to stop the spread while causing minimal loss.

1.5. INTEGRITY AND LONG-TERM VIABILITY

The ethical dilemma is that we can never guarantee a prediction to be perfectly accurate, and when the prediction is vague or inaccurate, it may send the wrong message to governments and their people, worsening the situation and causing greater loss. This highlights the need for continued research into COVID-19, which has the potential to save lives, reduce the loss, implement reasonable measures, and recover the global economy.

All of the data and modelling utilities used in the experiments are obtained online or from open-source repositories; no significant pollution is emitted; and the result is only beneficial to society for the reasons stated above, so the project poses no sustainability concern.

2. LITERATURE SURVEY

In this chapter, we give the most crucial information for readers to grasp the scope of this work: a brief introduction to the time-series analytic techniques employed in this research (ARIMA, Prophet, LSTM, and Transformer); a study of similar work; and a summary.

2.1. CHRONOLOGICAL DATA ANALYSIS

Statistics, economics, finance, and forecasting make extensive use of time-series data, which are collections of information that are indexed by timestamps or organised in time order. Time-series data are typically a sequence that has consistent intervals between points in time; for example, the data are taken every day at the same time.

The following time-series analysis techniques were selected because of their respective fields of competence. While each technique will be introduced later, it is currently unknown whether or not these techniques will continue to be employed and keep their outstanding track records.

In this research, we investigate whether or if these four methods can bring their strengths from the field of illness trend prediction to the novel field of forecasting the COVID-19 case trend, which we explain in detail below.

2.2. ARIMA

Auto Regressive Integrated Moving Average [5][6] is a popular time-series model in statistics and econometrics. It is a generalisation of the Autoregressive Moving Average (ARMA) model that incorporates differencing to address the limitation of the ARMA model to stationary time-series data.

Common notation for non-seasonal ARIMA includes the letters ARIMA (p, d, q), with definitions of each component provided as follows [7]:

- Autoregressive (AR) models are those in which the dependent variable is assumed to regress on its own lagged values, with the number of lags (the lag order) denoted by the parameter p.
- Time of differencing applied to raw data, denoted by d, causes time-series data to become stationary; this time is referred to as "integrated" (I).
- To show the relationship between an observation and the residual error of a moving average model applied to lagging observations, we use the notation MA, where the order of the moving average is denoted by q.

Typically, SARIMA models are written as SARIMA (p, d, q)(P, D, Q, m), where P, D, and Q are the same three components for the seasonal part of the model, and m is the number of periods in each season. The ARIMA models developed by Javier Contreras and colleagues have several uses. Rajesh G. Kavasseri and Krithika Seetharaman [8] analysed the hourly pricing data from the Spanish power markets using ARIMA models to estimate the next day's price with an average error of roughly 10%.

Predictions of wind speeds were made using f -ARIMA (a variant of ARIMA) [9]. It has been used to forecast diseases in recent years [10] and has improved accuracy by 42% compared to previous methods.

2.2. PROPHET

In 2017, the Core Data Science team at Facebook announced a forecasting method called Prophet [11], detailed in the paper "Forecasting at Scale" [12]. It's freely accessible online and may be downloaded in either Python or R. It is "based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects," as stated on the website.

Figure 1 shows an example of Prophet's prediction plot, where the black dots represent the observed values, the dark blue represents the predicted values, and the light blue represents the uncertainty. Prophet excels when the data have strong seasonal effects, handles missing data or data shifts in the series, and provides multiple user-friendly parameters to tune the model, adapting to specific domains to improve the performance.

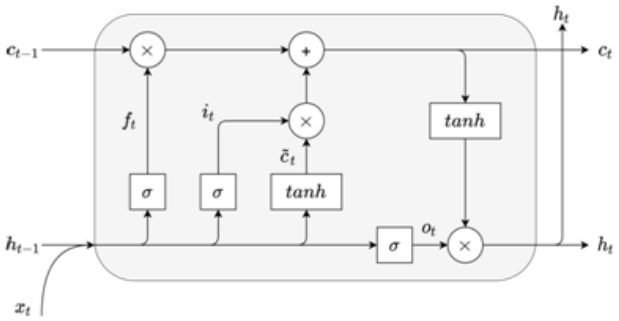

Fig. 1. Using Prophet, we may predict daily page views

Prophet's high accuracy in forecasting/prediction has led to its widespread use in a wide variety of applications across many different domains in recent years. For instance, [13] used Prophet to conduct Bitcoin forecasting, using Bitcoin values from May 3rd, 2016 to August 30th, 2018 as the training data to produce a 90-day forecast.

2.3. LSTM

LSTM was first proposed by [16] as a novel, efficient Recurrent Neural Network (RNN) architecture used in DL. It is created to address the vanishing gradient problem that traditional RNN can encounter when training. LSTM networks are applicable now in many different tasks, including handwriting recognition and speech recognition. It is overall well-suited for processing time-series data, because it captures the hidden information between events that are lagged with unknown intervals.

A typical LSTM unit consists of a cell, an input gate, an output gate, and a forget gate (shown in Figure 2), all of which can interact so that the unit receives short-term memory, long-term memory, and the input to generate new short-term memory, long-term memory, and the output. The input vector xt and the hidden state vector ht make up the LSTM unit's state space, while the cell state vectors ct and c t represent the forget gate, input gate, and output gate activation vectors, respectively, in the figure.

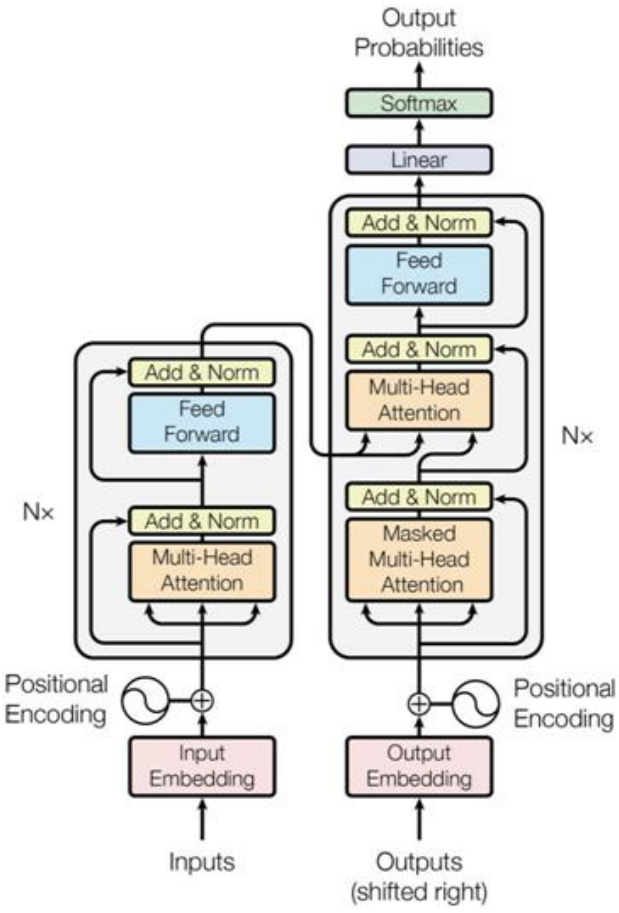

Fig .2. Structure of a Long Short-Term Memory (LSTM) unit

Over the course of two decades, researchers have studied LSTM and its variants, which have since found widespread use. For example, [17] used a bidirectional LSTM for sequence tagging, achieving state-of-the-art accuracy while the model relies less on word embedding. Similarly, [18] used convolutional LSTM as a novel approach to predict precipitation,

2.4. TRANSFORMER

The Transformer model, which uses the attention graph and was proposed in 2017 [20], has recently gained a lot of traction. Transformer is a model that uses a unique attention mechanism to process sequential data and allows for parallelization to reduce training time. Its main advantage over RNN is that it does not rely on the order in which data is processed.

Consisting, on the left, of an encoder stack and, on the right, of a decoder stack. Each encoder stack consists of a multi-head self-attention layer followed by a feed forward layer, while the decoder stack consists of a masked multi-head self-attention layer, a multi-head self-attention layer, and a feed forward layer.

Apart from the great achievement the Transformer has obtained in the fields of NLP and CV, it is also developed to adapt to time-series forecasting. [21] proposed Adversarial Sparse Transformer as a variation and proved its effectiveness and efficiency for both short-term and long-term prediction based on extensive experiments on real-world datasets. [22] used Transformer-based models to perform influenza-like illness (ILI) forecasting with the results favourably comparable to the state of the art.

2.5. SIMILAR RESEARCH

Related work is listed to provide an overview of the previous effort in these areas in light of the recent emergence of COVID-19. Researchers are focused on analysing the trend of the virus and making predictions about it using various methods and models.

[23] achieved 60-day forecast of COVID-19 using deep layer RNNs. In their study, the models analysed the 10 countries with the most confirmed cases, and for each country, a customised RNN model was proposed for better performance. They predicted the confirmed cases, recovered cases, and death cases in these countries, and they compared the accuracy of Gated Recurrent Units (GRUs) and LSTM units. Four DL models (LSTM, GRU, CNN, and Multivariate Convolutional Neural Networks) were used by [24] to study the COVID-19 epidemic in Brazil, Russia, and the United Kingdom.

However, LSTM was found to be inaccurate when depicting the trend because it attempted to capture the seasonality in the data, which is absent for the studied countries. Additionally, it is worth noting that a large amount of data is typically required for LSTM to capture the characteristics of data, which is not sufficient in their research.

[25] compared the accuracy of using ARIMA and Prophet to predict the spread of COVID-19 in Indonesia. They gathered data on confirmed cases, deaths, and recoveries beginning on March 2, 2020 and continuing for 81 days. They then used ARIMA and Prophet to predict cases in a 30-day forecast window. Results showed that both models achieved good performance in terms of forecast errors, but ANN is better than ARIMA in this study. [26] compared ARIMA and Artificial Neural Network (ANN) models to predict stock price. In this study, the closing price of the stock is chosen as the daily price to be modelled and predicted.

3. CONCLUSIONS

We gathered COVID-19 case data from JHU CSSE and identified six countries with the highest cumulative confirmed cases. We then evaluated five different time-series analysis methods, including a simple baseline model, to predict COVID-19 cases in these countries. The best-performing model achieved an accuracy of 0.06, with the Transformer model ranking second in accuracy in three of the countries, comparable to the LSTM model. The other models did not produce reliable results, and some even performed worse than the baseline model, particularly for India. Based on our analysis, we can conclude that the LSTM model remains a viable time-series analysis method for capturing disease trends, enabling us to forecast future trends and implement necessary measures.

REFERENCES

[1] H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino,E. Ortiz-Ospina, J. Hasell, B. Macdonald, D. Beltekian, and M. Roser, “Coronavirus Pandemic (COVID-19),” Our World in Data, Mar. 2020.

[2] L. Thunström, S. C. Newbold, D. Finnoff, M. Ashworth, and J. F. Shogren, “The Benefits and Costs of Using Social Distancing to Flatten the Curve for COVID-19,” Journal of Benefit-Cost Analysis, vol. 11, no. 2, pp. 179–195, 2020. doi: 10.1017/bca.2020.12 Publisher: Cambridge University Press.

[3] CSSEGISandData, “COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University,” Sep. 2021, original-date: 2020-02-04T22:03:53Z.

[4] V. Kotu and B. Deshpande, “Chapter 12 - Time Series Forecasting,” in Data Science (Second Edition), V. Kotu and B. Deshpande, Eds. Morgan Kaufmann, Jan. 2019, pp. 395–445. ISBN 978-0-12-814761-0.

[5] I. Yenidoğan, A. Çayir, O. Kozan, T. Dağ, and Arslan, "Bitcoin Forecasting Using ARIMA and PROPHET," in 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sep. 2018. doi: 10.1109/UBMK.2018.8566476 pp. 621–624

[6] G. A. Papacharalampous and H. Tyralis, "Evaluation of random forests and Prophet for daily streamflow forecasting," in Advances in Geosciences, vol. 45. Copernicus GmbH, Aug. 2018. doi: 10.5194/adgeo-45-201-2018 pp. 201–208, iSSN: 1680-7340.

[7] C. Xie, H. Wen, W. Yang, J. Cai, P. Zhang, R. Wu, M. Li, and S. Huang, "Trend analysis and forecast of daily reported incidence of hand, foot and mouth disease in Hubei, China by Prophet model," Scientific Reports, vol. 11, no. 1, p. 1445, Jan. 2021. doi: 10.1038/s41598-021-81100-2.

[8] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short- Term Residential Load Forecasting Based on LSTM Recurrent Neural Network," IEEE Transactions on Smart Grid, vol. 10, no. 1, pp. 841–851, Jan. 2019. doi: 10.1109/TSG.2017.2753802 Conference Name: IEEE Transactions on Smart Grid.

[9] S. Wu, X. Xiao, Q. Ding, P. Zhao, W. Ying, and J. Huang, "Adversarial Sparse Transformer for Time Series Forecasting," 2020

[10] N. Wu, B. Green, X. Ben, and S. O'Banion, "Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case," arXiv:2001.08317 [cs, stat], Jan. 2020, arXiv: 2001.08317. [Online]. Available: http://arxiv.org/abs/2001.08317

[11] K. E. ArunKumar, D. V. Kalaga, C. M. S. Kumar, M. Kawaji, and T. M. Brenza, "Forecasting of COVID-19 using deep layer Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) cells," Chaos, Solitons & Fractals, vol. 146, p. 110861, May 2021. doi: 10.1016/j.chaos.2021.110861. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S0960077921002149

[12] K. N. Nabi, M. T. Tahmid, A. Rafi, M. E. Kader, and M. A. Haider, "Forecasting COVID-19 cases: A comparative analysis between recurrent and convolutional neural networks," Results in Physics, vol. 24, p. 104137, May 2021. doi: 10.1016/j.rinp.2021.104137.

[13] C. B. Aditya Satrio, W. Darmawan, B. U. Nadia, and N. Hanafiah, "Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET," Procedia Computer Science, vol. 179, pp. 524–532, Jan. 2021. doi: 10.1016/j.procs.2021.01.036.

[14] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," The Lancet Infectious Diseases, vol. 20, no. 5, pp. 533–534, May 2020. doi: 10.1016/S1473-3099(20)30120-1 Publisher: Elsevier.

[15] "Time Series Made Easy in Python," Sep. 2021, original-date: 2018-09- 13T15:17:28Z.

[16] O'Malley, Tom, Bursztein, Elie, Long, James, Chollet, François, Jin, Haifeng, and Invernizzi, Luca, "KerasTuner," Sep. 2021, original-date: 2019-06-06T22:38:21Z.

[17] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A Transformer-based Framework for Multivariate Time Series Representation Learning," in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, ser. KDD '21. New York, NY, USA: Association for Computing Machinery, Aug. 2021. doi: 10.1145/3447548.3467401. ISBN 978-1-4503-8332-5 pp. 2114– 2124.

[18] Padhi, Y. Schiff, I. Melnyk, M. Rigotti, Y. Mroueh, P. Dognin, J. Ross, R. Nair, and E. Altman, "Tabular Transformers for Modeling Multivariate Time Series," in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Jun.2021. doi: 10.1109/ICASSP39728.2021.9414142 pp. 3565–3569, iSSN:2379-190X

[19] N. Harmony-search and otsu based system for coronavirus disease (COVID-19) detection using lung CT scan images. 2020 arXiv preprint arXiv:2004.03431.

[20] Bhapkar HR, Mahalle P, Dhotre PS. Virus graph and COVID-19 pandemic: a graph theory approach. Preprints 2020, 2020040507 (https://doi.org/10.20944/preprints202004.0507.v1).

[21] Bullock J, Pham KH, Lam CS, Luengo-Oroz M. Mapping the landscape of artifcial intelligence applications against COVID19. arXiv preprint arXiv:2003.11336. 2020.

[22] Mahalle PN, Sable NP, Mahalle NP, Shinde GR Data analytics: COVID-19 prediction using multimodal data. Preprints 2020, 2020040257 (https://doi.org/10.20944/preprints202004.0257.v1).

[23] Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U, Rabczuk T, Atkinson PM. Covid-19 outbreak prediction with machine learning. Available at SSRN 3580188. 2020.

[24] Santosh KC. AI-driven tools for coronavirus outbreak: need of active learning and cross-population train/test models on multitudinal/multimodal data. J Med Syst. 2020;44(5)1–5 https://doi. org/10.1007/s10916-020-01562-1.

[25] Dey N, Rajinikant V, Fong SJ, Kaiser MS, Mahmud M. Socialgroup-optimization assisted kapur's entropy and morphological segmentation for automated detection of COVID-19 infection from computed tomography images. 2020.

[26] Wagh CS, Mahalle PN, Wagh SJ. Epidemic peak for COVID19 in India, 2020. Preprints 2020, 2020050176 (https://doi. org/10.20944/preprints202005. 0176.v1).