Article | Received 6 August 2023; Accepted 10 August 2023; Published 1 December 2023 https://doi.org/10.55092/pcs2023020005

EEMD-LSTM modelling of daily confirmed COVID-19 cases in Malaysia

Ani Shabri¹, Siti Nabilah Syuhada Abdullah^{1,*}, Ruhaidah Samsudin², Zuriahati Mohd Yunos², Anita Fairos Ismail², Aida Ali²

- ¹ Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Malaysia
- ² School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Malaysia
- * Correspondence author; E-mail: nabilah1991@graduate.utm.my.

Abstract: The World Health Organization proclaimed COVID-19 to be in a pandemic state on March 11, 2020, when there were over 118000 confirmed cases worldwide across more than 110 countries. Accurate modeling and forecasting of the spread of confirmed and recovered COVID-19 cases are crucial for assisting decision-makers in fighting the epidemic. Such situations commonly exhibit non-linear patterns, motivating us to develop a system that can keep track of such alterations. The project's ultimate objective is to provide a method for anticipating new COVID 19 scenarios utilizing a hybrid EEMD-LSTM model. In this scenario, a prediction is produced regarding the total amount of daily COVID-19 cases that were officially confirmed in Malaysia between March 13, 2020, and January 4, 2021. The Global Change Data Lab at Oxford University provided the dataset.

Keywords: Ensemble Empirical Mode Decomposition (EEMD); Long-Short Term Memory (LSTM) network; forecasting COVID-19

1. Introduction

At the end of 2019, a lethal strain known as Corona-virus Disease 2019 (COVID-19) was found in Wuhan, China. Following the reporting of 118, 000 cases across 110 nations, on March 11th, 2020, the World Health Organization issued a pandemic announcement. Additional patient flows resulted in hospital bed shortages and high-stress circumstances around the country. It is essential to comprehend the trend and dissemination of this virus to support decision-makers [1-3]. Many modeling, estimation, and forecasting techniques are used to understand and control this epidemic.

Other publications assessed and predicted patterns in the COVID-19 outbreak in several nations, involving Italy, China, and India [4-9], using time-series strategies include ARIMA



Copyright©2023 by the authors. Published by ELSP. This work is licensed under Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

Shabri A, et al Proc. Comput. Sci. 2023(2):0005

(Auto-Regressive Integrated Moving Average) and Exponential Smoothing. Interestingly, machine learning, as a leading line of science for a wide range of reasons, has led these studies to [1,10-12] have given promising solutions to proactively improve disease growth prognostication. In Canada, Italy, and the US, COVID-19 dissemination was investigated using long short-term memory (LSTM) [12].

In accordance with best practice, two practice to representing time series data are used: the deterministic approach and the stochastic relationship between observations. The deterministic approach considers dynamic architectures [13], and statistical models have been constructed to support stochastic influences [14]. In reality, actually, time series data is a combination of both factors. In other cases, the model exactness is affected when one of the two processes is considered, *i.e.*, dynamical framework methods may produce distorted perturbations, while statistical models underestimate deterministic factors [15]. As a result of this limitation, our focus shifts to modelling time series based on deterministic and stochastic factors, employing various models for each component before combining the information for the final output with the goal of improving prediction performance. This situation necessitates a decomposition stage, which is accomplished in this study through the EEMD methodology [16].

The development of a method for COVID 19 case predictions is the main objective of this research. The data for this study came from the Global Change Data Lab at Oxford University and was gathered between March 13, 2020, and January 4, 2021. Section 2 would offer a thorough analysis of the tools and procedures used.

2. Description of data

Forecasting COVID 19 and predicting the spread of infection are the main objectives of this effort. This research is based on typical information from confirmed instances that were reported to Malaysia between March 13, 2020, and January 4, 2021. The Global Change Lab at Oxford University has made the data available. (https://ourworldindata.org/coronavirus) The COVID-19 A data set has been separated into a training set and a testing set in an 80:20 ratio to evaluate the model's competitiveness.

Figure 1 depicts the data graphically shown. Up until September 2022, this sickness has stayed well under control. Since its initial appearance, it has risen significantly, reaching a sizable total of 2525 verified cases on December 31, 2020. The data utilized in this analysis are described in Table 1. Here, we can infer that the dataset is non-Gaussian based on its Kurtosis value of more than 3, that the data is biased to the right of the positive Skewness value and that the data shows a wide variety of standard deviations.



Figure 1. Daily confirmed new cases of COVID-19 in Malaysia.

Min	1	
Max	2525	
STD	588.1898	
Q-0.25	16	
Q-0.5	404.9295	
Q-0.75	782.25	
Skew	1.531132	
Kurtosis	4.382176	

 Table 1. Summary of dataset.

3. Framework of study

The basic structure of the suggested methods for forecasting as seen in Figure 2. The time series was first broken down into several (k) Intrinsic Mode Functions (IMF) by the Ensemble Empirical Mode Decomposition. Premised on the autocorrelation graph, the partial graph, and the stationary Augmented Dickey-Fuller (ADF) graph, the best order of the ARIMA model is then specified for each IMF.

Each data strain that has attained p > 1 of the ARMA(p,q) is called a stochastic component and will therefore be modeled on an individual basis using the LSTM. After which merge these other deterministic IMFs that satisfy this requirement $p \le 1$. As a result, the list of IMFs required is now (k - n + 1), where n represents the total of IMFs to be merged as a Deterministic Variable. The information collected will then be standardized, and an LSTM model would be created. Here, Adam optimizer was pre-selected to reduce the loss function during testing. Subsequently, pre-constructed models with chosen parameters will be used to predict possible IMF values. Finally, the predicted parts are assembled to arrive at the predicted COVID-19 occurrences. The RMSE would be used to evaluate the accuracy of the developed model. RMSE will be used to assess the precision of the model developed.

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_t - y_t)^2}$$
(1)

In this study, predicting accuracy from the EEMD-LSTM, ARIMA, and standalone LSTM models will be compared to that from COVID-19 confirmed cases.



Figure 2. Framework of this study.

4. Data modelling

4.1. Ensemble Empirical Mode Decomposition (EEMD)

The method [17], facilitate data processing to accommodate data robustness. Characteristics of white noise in this method is that the maxima dispersion is temporally uniform at all timescales and is simply a two-channel tank for white noise [18-20]. Consequently, the degradation becomes more stable and physically significant.

Whilst using EEMD, the white noise magnitude and number of ensembles should be predetermined because they affect the decomposition process [21]. The EEMD technique used in this research is with a white noise magnitude of 0.2 times the standard divergence and a 100 ensemble. The work of [22] contains an in-depth examination of the ensemble amount and magnitude of noise.

The EEMD procedure was implemented to confirmed incidents of COVID-19 in Malaysia. The time series has been broken down into 7 IMFs and one residual. Figure 3 depicts the plots of all IMFs and residues that were autonomous.

As per Figure 3, IMF1, IMF2, and IMF3 have the maximum intensity biggest magnitude, and shortest wavelength. Following IMFs show the frequency and amplitude from highest to lowest, and vice versa for the wavelength. The penultimate residual variable provides the general pattern of the time series. The breakdown is extremely useful in converting non-linear

and non-stationary time series into stationary series to increase predictive power [22].



Figure 3. Decomposition of dataset using EEMD.

IMFs	ARIMA Model (p, d, q)
1	ARIMA (0, 0, 4)*
2	ARIMA (2, 0, 1)
3	ARIMA (2, 0, 0)
4	ARIMA (1, 0, 0)
5	ARIMA (0, 1, 0)*
6	ARIMA (0, 2, 0)*
7	ARIMA (0, 2, 0)*

Table 2. Criteria for reconstruction of IMFs.

*for p less than or equal to 1

Previously, EEMD divided the initial time series into strand known as IMFs. Following that, the best ARIMA model order for each strand was determined. Table 2 summarizes the results. The ARIMA model parameters p for the first, fifth, and subsequent IMFs are less than or equal to one, as shown in Table 2. The condition p 1 is now satisfied. As a result of the proposed legislation, one of these IMFs is being merged for further research. The total number of IMFs to be used for further research is then four.

4.2. Long-Short Term Memory (LSTM) network

Because it can acclimate to the non-linearity of the COVID-19 data set, the recurrent version of this network is a good forecasting candidate. Each LSTM block runs at a separate time

 (\mathbf{n})

(2)

step and passes the output to the following block before the last LSTM block provides a sequential output. Block RNNs (LSTMs) are analytic function for building a sequential time series model. Memory blocks, which have been configured to fix gradients decreasing by deciphering long-term network configuration, are the central aspect of LSTM networks. In digital technologies, the architecture is analogous to differential storage. Gates allow information to be managed using the activation function (Sigmoid) and output values ranging from 0 to 1. The sigmoid activation is used to achieve a straightforward result; we just need to transfer valid information to the next stage. The three gates of the LSTM network can be seen in the equations below:

$$J_t = sigmoid(w_I[h_{t-1}, k_t] + b_J$$
⁽²⁾

$$G_t = sigmoid(w_G[h_{t-1}, k_t] + b_G$$

$$P_t = sigmoid(w_p[h_{t-1}, k_t] + b_p$$
⁽⁴⁾

Where:

J_t = input gate operation	$k_t =$ input to the current operation at t time-
	step
$G_t = $ forget gate operation	h_{t-1} = output of previous time step
P_t = output gate operation	$w_x = coefficients of neurons at gate (x)$
	$b_x = bias of neurons at gate (x)$

In equation (2), the Input gate transmits messages to be stowed. The (3) equation coordinates the forget gate activation output, and the output of the forget gate at time step âtâ is then combined by the third equation to generate the throughput. Figure 4 depicts the underlying block diagram of the LSTM block implicated for the analysis. Self-loops were chosen to build a path in which gradients or weights can be traded over long periods of time. This is particularly useful when modelling profound learning models where the gradient of extinction is a common issue. Modifying the weights as self-locked gates would alter the time scale to track changing parameters. Figure 5 depicts the network topology used for this methodology.



Figure 4. LSTM internal architecture.



Figure 5. LSTM architecture.

Data from IMF 2, 3, 4 and Deterministic Component are fed into the LSTM model and forecasts are generated. The forecasted values are then summed up to produce the final forecast of confirmed new cases of COVID-19.

4.3. Performance of model

Figure 6 shows a comparison of forecasting capacity between the actual data (blue line), EEMD-LSTM (red line) model to basic ARIMA (grey line) and standalone LSTM model (yellow line). Table 3 on the other hand shows comparison of model between EEMD-LSTM, LSTM and ARIMA based on its RMSE. The EEMD-LSTM model was trained and tested on Malaysian dataset producing an RMSE error of 34.83 for short term predictions and about 45.70 for long term predictions. It is evident that EEMD-LSTM is the best model to be used in forecasting daily cases of COVID-19.



Figure 6. Modelling comparison of new COVID-19 cases in Malaysia.

	Training	Testing
EEMD-LSTM	34.83	45.7
LSTM	52.33	66.29
ARIMA	68.77	85.89

Table 3. Comparison of RMSE.

5. Conclusion

Though our model outperformed other analysis methods in terms of effectiveness, Unfortunately, the distribution is getting wider. In the meantime, both the incidence and total number of infections are expanding dramatically. If Malaysians fully shoulder the burden, the frequency of new instances will quickly start to reduce. These forecasts' accuracy is reliant on a variety of outside variables. To recapitulate, this is the first study to use data decomposition and machine learning methods to predict the intensity of COVID-19 in Malaysia.

Acknowledgement

This work was partially supported by UTMER grant, cost center 17J47. We would like to thank all parties involved from Universiti Teknologi Malaysia.

Conflicts of interest

The authors declare no conflict of interest.

References

- [1] Velásquez RM, Lara JV. Forecast and evaluation of COVID-19 spreading in USA with reduced-space Gaussian process regression. *Chaos Solitons Fractals*. 2020, 136:109924.
- [2] Yousaf M, Zahir S, Riaz M, Hussain SM, Shah K. Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan. *Chaos Solitons Fractals*. 2020, 138:109926.
- [3] Ribeiro MH, da Silva RG, Mariani VC, dos Santos Coelho L. Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. *Chaos Solitons Fractals*. 2020, 135:109853.
- [4] Dehesh T, Mardani-Fard HA, Dehesh P. Forecasting of covid-19 confirmed cases in different countries with arima models. *MedRxiv*. 2020:2020-03.
- [5] Gupta R, Pal SK. Trend Analysis and Forecasting of COVID-19 outbreak in India. *MedRxiv*. 2020:2020-03.
- [6] Chintalapudi N, Battineni G, Amenta F. COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: A data driven model approach. *J Microb Immunol Infect*. 2020, 53(3):396-403.
- [7] Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, Eggo RM, Sun F, Jit M, Munday JD, Davies N. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *Lancet Infect Dis.* 2020, 20(5):553-8.

- [8] Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet*. 2020, 395(10225):689-97.
- [9] Zhuang Z, Zhao S, Lin Q, Cao P, Lou Y, Yang L, He D. Preliminary estimation of the novel coronavirus disease (COVID-19) cases in Iran: A modelling analysis based on overseas cases and air travel data. *Int J Infect Dis.* 2020, 94:29-31.
- [10] Kırbaş İ, Sözen A, Tuncer AD, Kazancıoğlu FŞ. Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. *Chaos Solitons Fractals*. 2020, 138:110015.
- [11] Tuli S, Tuli S, Tuli R, Gill SS. Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Internet of Things*. 2020, 11:100222.
- [12] Chimmula VK, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos Solitons Fractals*. 2020, 135:109864.
- [13] Alligood KT, Sauer TD, Yorke JA. One-dimensional maps. In *Chaos: An Introduction* to Dynamical Systems. 1996:1-42.
- [14] Box GE, Jenkins GM, Reinsel GC, Ljung GM. *Time Series Analysis. Forecasting and Control.* John Wiley & Sons, Hoboken. 2015.
- [15] Han M, Liu Y. Noise reduction method for chaotic signals based on dual-wavelet and spatial correlation. *Expert Syst Appl.* 2009, 36(6):10060-7.
- [16] Aamir M, Shabri A, Ishaq M. Improving forecasting accuracy of crude oil prices using decomposition ensemble model with reconstruction of IMFs based on ARIMA model. *Malays J Fundam Appl Sci.* 2018, 14(4):471-83.
- [17] Huang NE, Shen Z, Long SR. A new view of nonlinear water waves: the Hilbert spectrum. *Annu Rev Fluid Mech.* 1999, 31(1):417-57.
- [18] Flandrin P, Rilling G, Goncalves P. Empirical mode decomposition as a filter bank. *IEEE Signal Process Lett.* 2004, 11(2):112-4.
- [19] Wu Z, Huang NE. A study of the characteristics of white noise using the empirical mode decomposition method. In *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences.* 2004, 460(2046): 1597-611.
- [20] Wu Z, Huang NE. On the filtering properties of the empirical mode decomposition. *Adv Adapt Data Anal.* 2010, 2(04): 397-414.
- [21] Wu Z, Huang NE. Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv Adapt Data Anal*. 2009, 1(01):1-41.
- [22] Debert S, Pachebat M, Valeau V, Gervais Y. Ensemble-empirical-mode-decomposition method for instantaneous spatial-multi-scale decomposition of wall-pressure fluctuations under a turbulent flow. *Exp Fluids*. 2011 Feb, 50:339-50.