Journal of Theoretical and Applied Information Technology

15th October 2023. Vol.101. No 19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DATA MINING FOR COVID-19 NEW CASES AND DEATH FORECASTING IN INDONESIA

FREDDY KURNIAWAN SOEBARKAH[1] , LILI AYU WULANDHARI[2]

[1] Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

[2] Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

E-mail: [1] freddy.soebarkah@binus.ac.id, [2] lili.wulandhari@binus.ac.id

ABSTRACT

COVID-19 has been a major threat to Indonesians and the world in many life aspects. Therefore, it is important to predict the covid-19's new cases and death accurately to anticipate the rise of covid-19 cases. The goal of the research is to develop the time-series prediction model to predict the number of Indonesia's COVID-19 new cases and death. In this research, we conduct to cluster the locations of data before the prediction process. The creation of clustering model is done as the early step before COVID-19 prediction because of so many data variations exists across the 34 provinces. We use K-Means Dynamic Time Warping (DTW) method to cluster the provinces and some comparison in machine learning and deep learning approach for prediction model as much as the number of clusters. The LSTM is chosen as the deep learning approach where we compare between the benchmark from previous research and our proposed model, together with Support Vector Regressor (SVR) as the machine learning approach. Indonesia's public COVID-19 data with periods from 1 March 2020 to 3 December 2021 is used for training and testing the model. The experiment results show the best number of clusters is three, and from RMSE and MAE, our new cases and new deaths model have lower error and less overfit. The proposed model improved 29.11% higher for RMSE and 42.55% for MAE respectively than the SVR. While it achieves 20.22% improvement for RMSE and 16.24% for MAE respectively than the state-of-the-art LSTM.

Keywords: *COVID-19, Prediction, Time series, Long Short Term Memory, Data Mining*

1. INTRODUCTION

COVID-19 has been a major threats to the world in many life aspects [1].When Indonesia confirms its first case of COVID-19 at March 2020, the weak health control in Indonesia leads to the massive spread of COVID-19 in Indonesia and caused the crisis in economics, health, and political aspects. This eventually forces the Indonesia's government to put a high priority on handling the economy in the face of the COVID- 19 viruses, known as the 'new normal' policy which was established in June 2020 [2]. Knowing the large impact of COVID-19, according to Central Disease Centre (CDC), prediction tools will be very helpful in predicting the number of transmissions, deaths, and hospitalizations. Several studies have been conducted to predict the number of COVID-19 with a machine learning approach. Muhammad, *et al*. compared several models in the form of Decision Tree, Random Forest, Logistic Regression Random Forest, and K-Nearest Neighbor using python to predict the number of cases COVID-19 with Decision Tree results that have the best performance [3]. Ayyoubzadeh, *et al*. developed a Long Short Term Memory (LSTM) to predict covid time series data. The result is a Root Mean Square Error (RMSE) of 7.562 for Linear Regression and 27.187 for LSTM [4]. Ballı, S. developed Random Forest, Linear Regression, Multilayer Perceptron, and Support Vector Machine (SVM) to predict time-series data of COVID-19 cases in Germany, America, and the world. As a result, SVM gives the best prediction results with the lowest Root Mean Square Error (RMSE), Absolute Percentage Error (APE), and Mean Absolute Percentage Error (MAPE) [5]. From these studies, the amount of research to predict COVID-19 cases in Indonesia is very rare, as evidenced by most of the predicted COVID-19 data are data from western countries, like USA, the European continent, and there is one from Iran. Knowing the difference in policies for each country (including Indonesia) in dealing with COVID-19 which will also affect the pattern of the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

number of COVID-19 cases, it is important to be able to look at the prediction model for the number of COVID-19 in Indonesia so that the best prediction model can be used in Indonesia in dealing with the pandemic. Of course, we need to use a model that works best in handling time series data. Ayyoubzadeh, *et al*. said that LSTM works best in predicting time series data [4], [6] because LSTM stores the previous prediction results to predict the data in the next timestep so that it can provide accurate prediction results. Therefore, the implementation of this research is to use the development of the LSTM model. The LSTM models will be compared with SVR and LSTM from Ayyoubzadeh, *et al.* [4] models. In this research, the prediction covers for New Cases and New Deaths for 34 provinces in Indonesia. We did not cover New Active and New Recovery because New Active information can be obtained as the summary of New Cases, meanwhile New Recovery is not covered because the data is not updated as the patients not always informed when they are recovered. Knowing the huge differences between each 34 provinces, rather than making all 34 models for each LSTM, SVR, and LSTM Ayyoubzadeh, *et al.* [4] for new cases and another 34 models for new deaths, we decided to divide the data to several clusters named K to see the similar data, and then we can just make K amount of models for each LSTM, SVR, and state-of-the-art LSTM for new cases and another K amount for new deaths..

2. MATERIALS AND METHOD

2.1 Related Works

Our related works are related to prediction, clustering, and COVID-19. The related works about COVID-19 cases prediction has been mentioned in the Introduction. We determine our algorithm for this research from the related works.

Some of clustering research has been done by previous researchers some of them are Sinaga *et al.* improved a k-means algorithm so we don't need to manually define the number of clusters or k [7]. Song *et al.* adopted a decision tree to cluster cyberbullying from social medias, with results there are victims, perpetrators, and bystanders [8]. Abdullah *et al*, uses k-means to clusters province in Indonesia in terms of COVID-19, the results is there are 3 clusters [9]. T. Li *et al,* develop an algorithm using several Agglomerative Hierarchical Clustering (AHC) method to combat ensemble clustering problem on big data [10]. M. Zhang, *et al.* improves k-means time series clustering to predict the intensity of solar radiation [11]. Teichgraeber, *et al.* compares k-means, hierarchical clustering, k-medoids, k-shape, and dynamic time warping. Centroid based (K-means) provides best result in gas turbine scheduling and battery charge/discharge optimization [12]. Li, *et al.* uses K-Means Dynamic Time Warping to cluster bike sales pattern from time series data [13]. Mashtalir, *et al.* use matrix harmonic k-means to do segmentation-clustering from video sequences [14]. Abbas, *et al.* uses k-medoids and k-means to cluster the birth data in Kashmir. The k-medoids performed better here [15].

After clustering, we analyze the research about prediction using regression and deep learning.

Khare, *et al.* used Linear Regression, Polynomial Regression, Decision Trees Regression, and Random Forest Regression to predict real estate cost. The Polynomial Regression performed the best [16]. G. Fan, *et al.* combines Support Vector Regressor (SVR) with Grey Catastrophe, and Random Forest to forecast electricity load from time series data [17]. Johannesen, *et al.* uses Random Forest regressor, k-Nearest Neighbour regressor, and linear regressor to predict electrical demand from time series data. Random Forest is the best for short-term prediction and k-nearest neighbour is the best for long-term prediction [18]. Hosseini, *et al.* use Multiple Polynomial Regression and Multiple Linear Regression to forecast CO_2 emission in Iran based on time series data [19]. Xu, *et al*. combines Linear Regression and Deep Belief Network to predict a time series data [20]. Z. Zhang, *et al*. combines Support Vector Regressor with the chaotic mapping mechanism, variational mode decomposition, and the grey wolf optimizer to forecast electric load from time series data [21]. Maldonado, *et al.* uses Support Vector Regressor with Gradient Descent to forecast electric load from time series data [22]. Xing, *et al.* uses LSTM based neural network to forecast wireless traffic from fluctuation time series data [23]. Bilgili et al. tested LSTM, adaptive neuro-fuzzy inference system (ANFIS) with Subtractive Clustering (SC), ANFIS with Grid Partition (GP), and ANFIS with fuzzy cmeans (FCM) to predict electrical energy consumption via time series data. The evaluation was carried out with Correlation Coefficient (R), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). LSTM has the best performance here [24]. Park, *et al.* uses LSTM to predict the battery useful life remaining with multi-channel charging through time series data [6].

From the related works, we decided to analyze and determine the machine learning for this

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

research by analyzing the pros and cons of each of the algorithms with best results from each paper. First from the clustering algorithm and then the prediction algorithm. The italicized pros are the strong reason why we decided to use the algorithm.

Table 1: Pros and Cons for the Clustering Algorithm

Clustering Algorithm	Pros	Cons
K-Means Dynamic Time Warping	• Fast convergence • *Super effective in clustering time series data* • *The most promising approach in time series clustering*	• Too much variation and density of the data can cause problems in k-means
Autoregressive	• Meant to be used in forecasting. • Has great accuracy	Uses less prior information which leads to exclusions of the important variables.
AHC	• Easy to use. • No need to define cluster number	• High complexity

Table 2: Pros and Cons for the Prediction Algorithm

Prediction Algorithm	Pros	Cons
Support Vector Regressor	• Fast runtime • *Suitable to be used on data with a lot of features*	• Overfits are difficult to be detected and fixed. • Not suitable for large datasets
Long Short Term Memory	• *Highly suitable for time series prediction* • Overfitting can be detected and fixed easily. • *Provides high accuracy across papers*	• Requires a lot of resources including runtime. • Cannot remove vanishing gradient problem completely
Decision Tree	• Easy to handle when overfits. • No need so much data preparation like scaling • Not affected by outliers	•Not suitable to be used on continuous numerical data time series data
Random forest	• Not affected by outliers • Automatically selects important features	•Not suitable for large datasets •We cannot tune the model in case of bad performance
Adaptive Neural Fuzzy Inference System (ANFIS)	• Capture nonlinear structure. • Able to adapt to any datasets. • Fast learning	•Not suitable datasets with a lot of features •Slow runtime •Lose to LSTM in terms of performance
Gradient Boosting Machine	• Fast training time on large datasets • More accurate than random forest	•Slow runtime •Hard to interpret the final model

2.2 Dataset

We used Indonesia COVID-19's public data from Kaggle (https://www.kaggle.com/hendratno/covid19-indonesia/version/84) ranging from the 1st of March 2020 until 3rd December 2021. The data denotes COVID-19 cases daily in all 34 provinces in Indonesia, so there are 642 days per province in total. The data variables are: 'Date', 'Location', 'New Cases', 'New Deaths', 'New Recovered', 'New Active Cases', 'Total Regencies', 'Total Cities', 'Total Districts', 'Total Urban Villages', 'Total Rural Villages', 'Area (km^2)', 'Population', 'Population Density', 'Longitude', 'Latitude', 'Growth Factor of New Cases', 'Growth Factor of New Deaths', 'Total New Cases', 'Total New Deaths', 'Total New Recovered', 'Total Active Cases', 'New Cases per Million', 'Total Cases per Million', 'New Deaths per Million', 'Total Deaths per Million', 'Total Deaths per 100rb', 'Case Fatality Rate', 'Case Recovered', 'Special Status', 'Continent', 'Country', 'Time Zone', 'Province', 'Location Level', 'Island', 'Location ISO Code', and 'City or Regency'. From this dataset, the variable we want to predict are 'New Cases' for new cases prediction, and 'New Deaths' for new

deaths prediction. 'New Cases' will be omitted from the feature lists for 'New Deaths' prediction and 'New Deaths' will be omitted for 'New Cases' prediction. For the features, we going to use all data variables except the aggregate variables ('Total Cases', 'Total Deaths', 'Total New Recovered', 'Total Active Cases', 'New Cases per Million', 'Total New Cases per Million', 'New Deaths per Million', 'Total New Deaths per Million', and 'Total New Deaths per 100rb'), variables with value more than 100% ('Case Fatality Rate' and 'Case Recovered'), and non-numeric variables ('Special Status', 'Continent', 'Country', 'Time Zone', 'Province', 'Location Level', 'Island', 'Location ISO Code' ,'City or Regency'). The 'Location' is known by sorting the data by 'Dates' and 'Location'. So, there are 642 days for each 'Location'. Finally, we made sure that there are 642 data for each location/province.

2.3 Theories

K-means clustering is useful for grouping data into several clusters as many as K from a data set. This algorithm works by determining the initial cluster center which is done by randomly selecting K amount of cluster centers. The next step is to use Lloyd's Iteration whose steps are: 1) Calculate the Euclidean distance stated in equation (1) between each data x and y coordinates to a predefined cluster centers and then assigns this coordinate to the nearest cluster center. Below is the formula of Euclidean distance:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (1)$$

p and q is two different data point, n is the amount of data, and i is the ith data.

2) Update the cluster center by calculating the mean of the data point values corresponding to the data coordinates in each cluster. 3) Calculates the Euclidean distance in equation (1) between the cluster centers of two iterations. The iteration is stopped when the distance is lower than a certain condition/threshold like epoch, otherwise the iteration is continued [25], [26].

Elbow method is an algorithm used to determine the best K for K-Means by counting the inertia of the Sum of Squared Error which is:

$$SSE = \sum_{K=1}^{K}\sum_{x_i \in S_k}\|X_i - C_k\|^2 \quad (2)$$

X_i is the ith data, C_k is the *cluster center* from cluster K. The optimum K is determined from the altered SSE graph from downturn into linear shaped line.[26]

Silhouette method is also used to determine the optimum K in K-Means. Silhouette get the optimum K from the highest Silhouette Index (*SI*) whose formula is:

$$SI = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (3)$$

b_i is the smallest Euclidean distance mean from the ith data to the other data, a_i is the Euclidean distance mean from the ith data to the other data. $\max(a_i, b_i)$ is the bigger value between a_i and b_i. A good SI is between 0.7 until 1.0 [26]

Dynamic Time Warping is an algorithm that can measure Euclidean distance between two time series data even when the two time series data do not have the same length of data series. The warping path $w_k = (x_i(k), y_i(k))$ between 2 time series is displayed by the mapped elements of the time series into a $m \times n$ sized matrix. This warping path is the distance between the w_k coordinates indicating the sequence between the $x_i(k)$ and $y_i(k)$ coordinates of the time series. The length L of this warping path must be between m and n, or m L n assuming n m. DTW chooses a 'warping path' for two time series data which gives the minimum Euclidean distance between each coordinate. Below is the formula of DTW:

$$DTW(TS1, TS2) = \min(\sum_{k=1}^{L} d(w_k)) \quad (4)$$

Chen, *et al.* eventually uses DTW because this algorithm provides the most promising approach [27], [28].

MinMaxScaler is a famous method to rescale the data, so the data range becomes 0 until 1. This rescale is very useful to ease the data processing. The formula is :

$$v' = \frac{v - \min(v)}{\max(v) - \min(v)} \quad (5)$$

v' is the normalized data, v is the unnormalized data. $\min(v)$ is the minimum value in v, $\min(v)$ is the maximum value in v [29].

After the clustering, we did the prediction uses LSTM and SVR. LSTM is a Recurrent Neural Network (RNN) model. It was made to deal with vanishing gradient problem which disrupt the model's performance in common RNN models by adding memory layers with constant error. LSTM layers are managed by 3 gates: input, output, and forget. Training process is maintained until maximum epoch or minimum error is reached.

LSTM is very effective in predicting time series data because LSTM works by saving the result of the previous prediction in short term and long term memory, so it can predict more accurately.

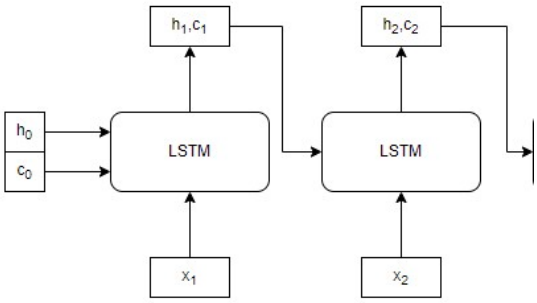

Figure 1: LSTM Architecture

In fig. 1, $x = (x_1, x_2, \ldots, x_n)$ is the input (time series data) used to get the cell state c which is $c = (c_1, c_2, \ldots, c_n)$ and hidden state which is $h = (h_1, h_2, \ldots, h_n)$. x_1 is used to get the first updated cell state (c_1) and first updated hidden state (h_1) at the first LSTM unit. At time step t, h_{t-1} and c_{t-1} is inputted to the first LSTM unit to get h_t and c_t . Hidden state (h_t) at time step t is calculated by:

$$h_t = o_t \odot \tanh(C_t) \quad (6)$$

$\odot$ is multiplicative vector named Hadamard product. o_t is the output gate which manages the connected cell state with the hidden state. Cell state adds or removes information from LSTM unit to control the LSTM network. At time step t, cell state (c_t) incorporates information gathered from the previous LSTM unit by:

$$c_t = f_t \odot c_t - 1 + i_t \odot g_t \quad (7)$$

Forget gate (f_t) directs the resets' degree of the cell state, and input gate (i_t) manages the update of the cell state. Candidate cell (g_t) adds information to the cell state. These processes are done with:

$$i_t = \sigma(W_i x_t + R_i h_{t-1} + b_i) \quad (8)$$

$$f_t = \sigma(W_f x_t + R_f h_{t-1} + b_f) \quad (9)$$

$$g_t = tanh(W_g x_t + R_g h_{t-1} + b_g) \quad (10)$$

$$o_t = \sigma(W_o x_t + R_o h_{t-1} + b_o) \quad (11)$$

σ is sigmoid function which is: $\sigma(x) = (1 + e^{(-x)})^{-1}$. R are recurrent weights with the details: R_i = recurrent weight at input gate, R_g = recurrent weight at candidate cell gate, R_f = recurrent weight at forget gate, and R_o = recurrent weight at output gate. W is input weights with details : W_i = input weight at input gate, W_g = input weight at candidate cell gate, W_f = input weight at forget gate, and W_o = input weight at output gate and b is bias with details : b_i = bias at input gate, b_g = bias at candidate cell gate, b_f = bias at forget gate, and b_o = bias at output gate [4], [24].

Support Vector Regressor (SVR) is part of the Support Vector Machine (SVM) in the form of a model that predicts Regression using a linear function in a large scope. This SVR maps features or data other than 'New Cases' for prediction models for the number of new cases and data other than 'New Deaths' for mortality prediction models into high-dimensional data with non-linear transformations [5], [30].

Root Mean Square Error (RMSE) is an evaluation method to measure the performance of a machine learning model[31]. RMSE is calculated by the formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (12)$$

y is the predicted result, $\hat{y}$ is the actual result, n is the amount of data, and i is the ith data.

Mean Absolute Error is an evaluation method to measure the performance of a machine learning model [31]. Mean Absolute Error (MAE) is calculated by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (13)$$

y is the predicted result, $\hat{y}$ is the actual result, n is the amount of data, and i is the ith data.

2.4 Proposed Method

In this research, we use clustering and time series prediction which is mentioned in Figure 2.

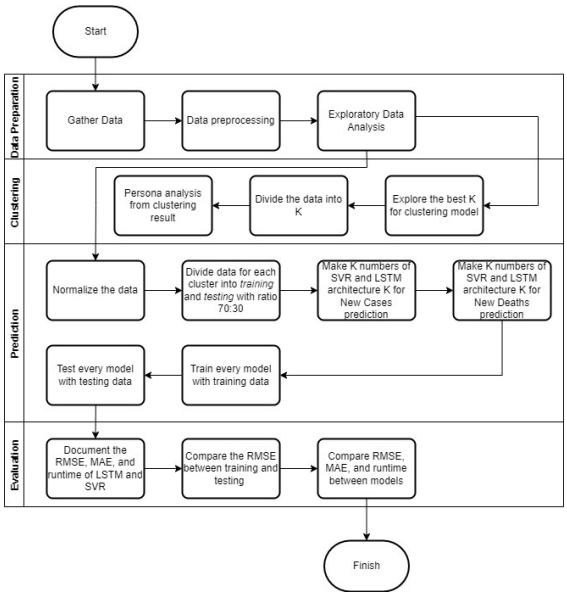

Figure 2: Research Methodology

We gather the data from Kaggle, then we removed the aggregate, non-numeric, and confusing variables (value more than 100%). Then we analyze the data through exploratory data analysis to recognize the COVID-19 new cases and new deaths pattern, also to see the important variable. We use K-Means Dynamic Time Warping to cluster the data into K clusters. To obtain the best K value, we use the Elbow and Silhouette methods with K values tested from 1 to 34 (number of provinces in Indonesia). The cluster was then analyzed to see the characteristic of each cluster. Our training and testing ratio will be 70:30 for each cluster. which means training data consist of 450 days and testing data consist of 192 days. Before we use the data in the prediction model, we normalized the data with MinMaxScaler.

For the prediction model, we were improving the LSTM model developed by Ayyoubzadeh, *et al.* [4]. The reason was that although Linear Regression has a smaller RMSE value, the LSTM works by storing the results from the previous prediction data, so it is more suitable to be used in predicting time series data. Therefore, the LSTM development will be more suitable. The LSTM will be trained and evaluated repeatedly until the model is not overfit. We will develop K amount of LSTM for New Cases prediction, and another K amount of LSTM for New Deaths prediction. The number 14 in Figure. 3 is based on the number of features (X coordinate) being used in the model.

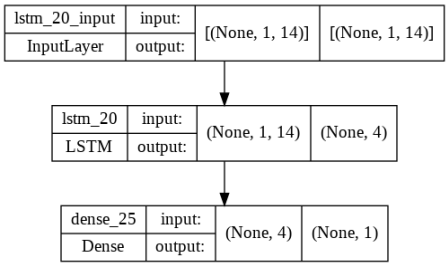

Figure 3: The LSTM development design

For the comparison model, we used the SVR. This SVR will be hyperparameter tuned with the training data. We will develop K amount of SVR for New Cases prediction, and another K amount of SVR for New Deaths prediction.

For the evaluation, we compared the RMSE and MAE between our LSTM and SVR and between our LSTM and LSTM model developed by Ayyoubzadeh, *et al.* [4], to see the improvement of our model. We also compare the training vs validation of the to see whether the model is overfit or not.

3. RESULT AND DISCUSSION

3.1 Clustering Result and Analysis

From $k = 1$ until $k = 33$, the elbow graph curved the most when $k = 3$ which means 3 is the best k in fig. 4. From $k = 2$ until $k = 33$, silhouette index shows 0.71 at $k = 3$ in fig. 5 which means the data is clustered well when $k = 3$. This means the data is best to be clustered by 3.

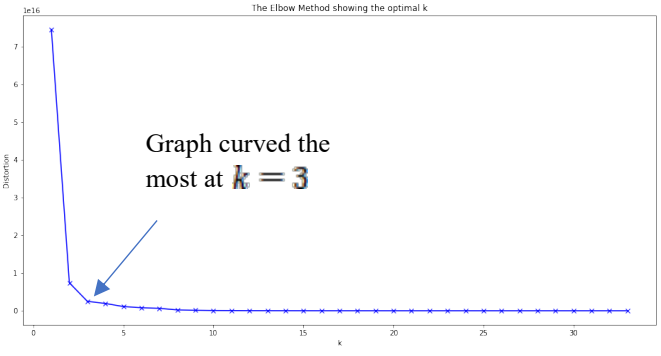

Figure 4: Elbow Graph

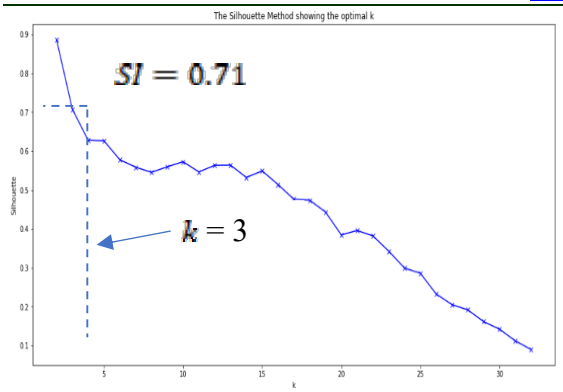

Figure 5: Silhouette Graph

Table 3 shows the cluster analysis by analyzing the comparisons of the new cases and population, the population density, the population, rural and urban areas, and new deaths and population between these 3 clusters. Basically, we analyze the environmental factors and how the covid was treated in each cluster.

3.2 Training and Validation Comparison

We compares the training and validation RMSE between our LSTM and LSTM developed by Ayyoubzadeh, *et al.* [4] to see which one is more overfit.

Figure 7 shows our LSTM is less overfit, shown by lower Validation Loss than LSTM Ayyoubzadeh *et al.* [4] in Figure 6.

Table 3: Clusters Persona Analysis

	Cluster 0	Cluster 1	Cluster 2	Analysis
Comparison of new cases and population	0.0020%	0.0034%	0.0029%	Cluster 1 has the largest percentage of spread
Comparison of death rate and population	0.00009%	0.00006%	0.00007%	Cluster 0 has the largest percentage of deaths
Comparison of total new cases and cure rates	95%	98%	96%	
Comparison of modern areas with rural areas	9%	16%	8%	
Equator (Equator = 0)	-7.30096	-3.72227	-1.96285	COVID-19 spreads the fastest in Cluster 0, because Cluster 0 tends to be colder[32]
Average population	40479000	10074600	2798730	Cluster 0 factor has the largest number of new cases and deaths
Population density (per km^2)	1077	3033	156	Cluster 1 factor has the largest percentage of new cases

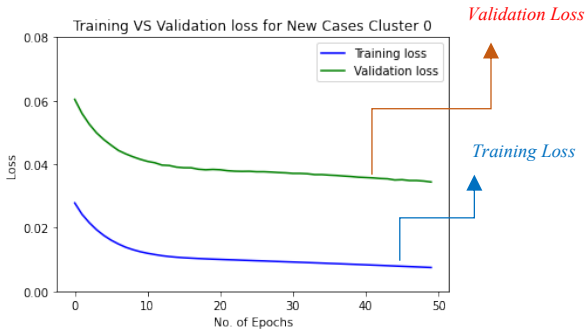

Figure 6: Training and Validation New Cases Cluster 0 of LSTM Ayyoubzadeh

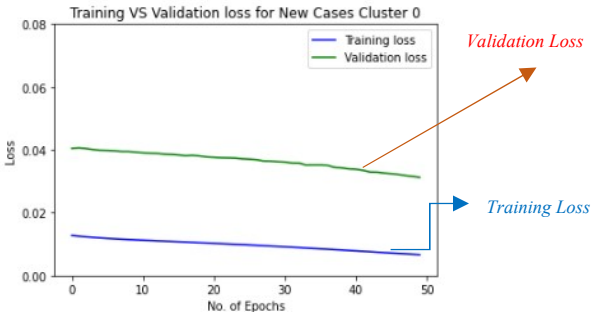

Figure 7: Training and Validation New Cases Cluster 0 of our LSTM (bottom)

Figure 9 LSTM shows our LSTM is less overfit, shown by lower Validation Loss than LSTM Ayyoubzadeh, *et al*. [4] in Figure 8.

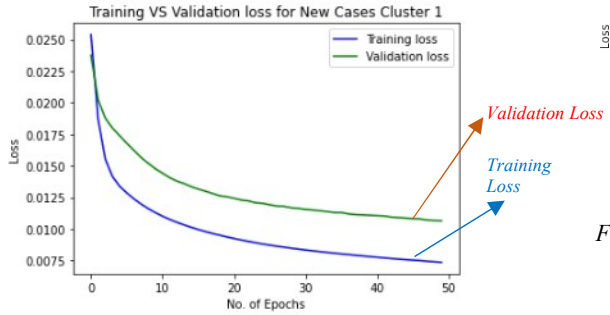

Figure 8: Training and Validation New Cases Cluster 1 of LSTM Ayyoubzadeh

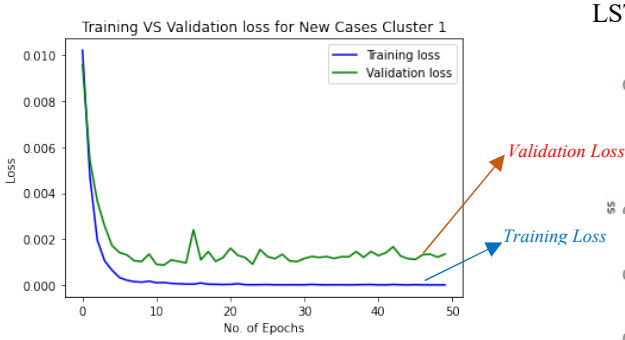

Figure 9: Training and Validation New Cases Cluster 1 of our LSTM

Figure 11 LSTM shows our LSTM has the same Validation Loss with LSTM Ayyoubzadeh, *et al.* [4] in figure 10 which means they are the same.

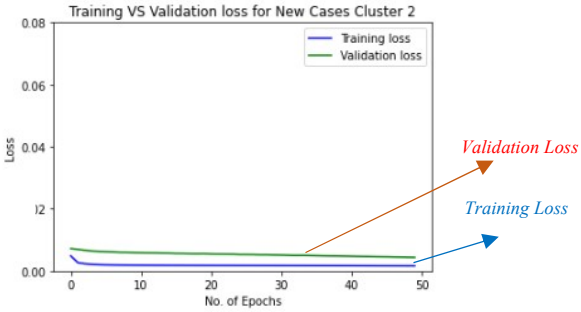

Figure 10: Training and Validation New Cases Cluster 2 of LSTM Ayyoubzadeh

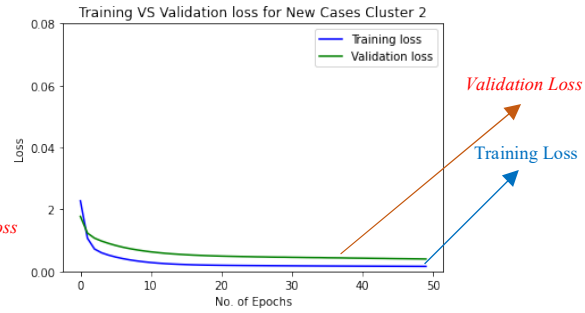

Figure 11: Training and Validation New Cases Cluster 2 of our LSTM

Figure 13 LSTM shows our LSTM is less overfit, shown by lower Validation Loss than LSTM Ayyoubzadeh, *et al.* [4] in Figure 12.

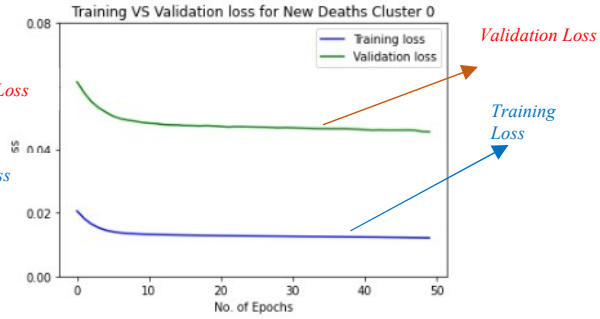

Figure 12: Training and Validation New Deaths Cluster 0 of LSTM Ayyoubzadeh

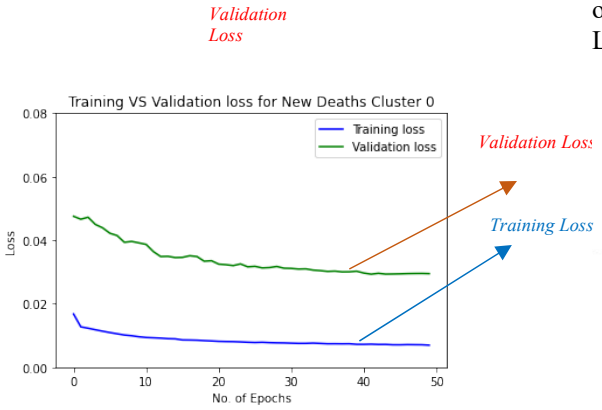

Figure 13: Training and Validation New Deaths Cluster 0 of our LSTM

Figure 15 LSTM shows our LSTM is less overfit, shown by lower Validation Loss than LSTM Ayyoubzadeh, *et al.* [4] in figure 14 .

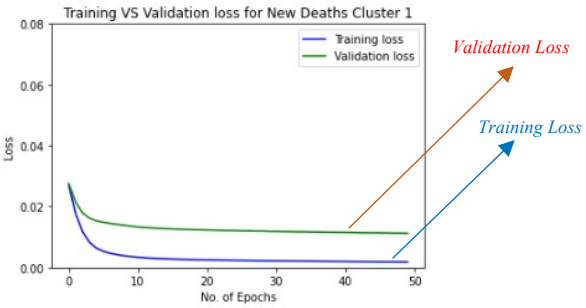

Figure 14: Training and Validation New Deaths Cluster 1 of LSTM Ayyoubzadeh

Figure 15: Training and Validation New Deaths Cluster 1 of our LSTM

Figure 17 LSTM shows our LSTM is less overfit, shown by lower Validation Loss than LSTM Ayyoubzadeh *et al.* [4] in figure 16 .

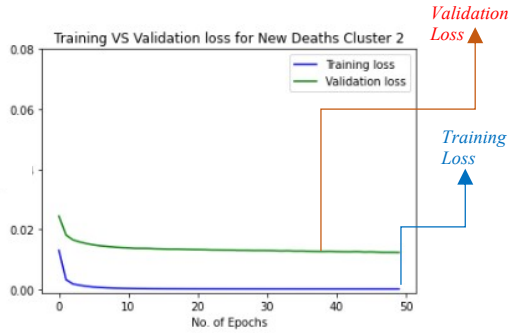

Figure 16: Training and Validation New Deaths Cluster 1 of LSTM Ayyoubzadeh

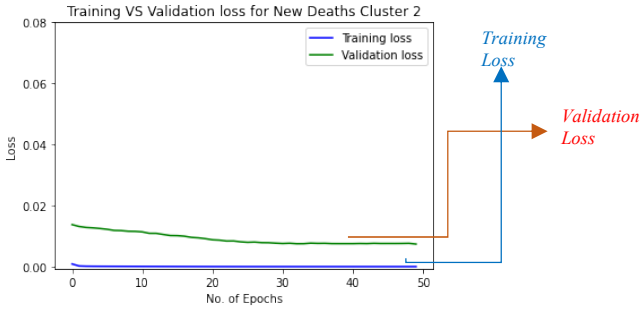

Figure 17: Training and Validation New Deaths Cluster 2 of our LSTM

From Training and Validation comparison, our 5 over 6 LSTM has a lower distance between the training and validation graph and the other 1 of our LSTM has roughly the same training and validation graph. This means our LSTM is less overfit than the LSTM Ayyoubzadeh, *et al.*[4]

3.3 RMSE Comparison

We compares the RMSE of our LSTM with the LSTM developed by Ayyoubzadeh, *et al.* [4] in table 4 and with SVR in table 5 as well as our LSTM improvement value. The marked table indicates our LSTM has better performance.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4: Our LSTM and LSTM Ayyoubzadeh Comparison in RMSE

RMSE				
Model		**LSTM**	**LSTM** [4]	**Improvement (%)**
Cluster 0	New Cases	**0.186**	*0.177*	-5.08%
	New Deaths	**0.172**	*0.208*	17.30%
Cluster 1	New Cases	**0.037**	*0.103*	64.07%
	New Deaths	**0.087**	*0.106*	17.92%
Cluster 2	New Cases	**0.063**	*0.066*	4.54%
	New Deaths	**0.086**	*0.111*	22.52%

The Improvement (%) was gathered by counting this formula in table 4:

$$\frac{(LSTM - LSTM\ Ayyoubzadeh)}{LSTM\ Ayyoubzadeh} \times 100\% \quad (14)$$

The average of the Improvement (%) from table 4 or LSTM Ayyoubzadeh, *et al.* [4] is 20.22 %.

Table 5: Our LSTM and SVR Comparison in RMSE

RMSE				
Model		**LSTM**	**SVR**	**Improvement (%)**
Cluster 0	New Cases	**0.186**	*0.198*	6.06%
	New Deaths	**0.172**	*0.204*	15.68%
Cluster 1	New Cases	**0.037**	*0.089*	58.42%
	New Deaths	**0.087**	*0.083*	-4.81%
Cluster 2	New Cases	**0.063**	*0.201*	68,65%
	New Deaths	**0.086**	*0.124*	30.64%

The Improvement (%) was gathered by counting this formula on table 5:

$$\frac{(LSTM - SVR)}{SVR} \times 100\% \quad (15)$$

The average of the Improvement (%) from table 5 or SVR is 29.11 %.

3.4 MAE Comparison

We compares the MAE of our LSTM with the LSTM developed by Ayyoubzadeh, *et al.* [4] in table 6 and with SVR in table 7 as well as our LSTM improvement value. The marked table indicates our LSTM has better performance.

Table 6: Our LSTM and LSTM Ayyoubzadeh Comparison in MAE

MAE				
Model		**LSTM**	**LSTM.** [4]	**Improvement (%)**
Cluster 0	New Cases	**0.128**	*0.118*	-8,474%
	New Deaths	**0.129**	*0.165*	21,818 %
Cluster 1	New Cases	**0.033**	*0.057*	42,105 %
	New Deaths	**0.039**	*0.050*	22%
Cluster 2	New Cases	**0.034**	*0.034*	0%
	New Deaths	**0.036**	*0.045*	20%

The Improvement (%) was gathered by counting equation (14) on table 6. The Improvement (%) average from table 6 or LSTM Ayyoubzadeh, *et al.* [4] is 16.24 %.

Table 7: Our LSTM and SVR Comparison in MAE

MAE				
Model		**LSTM**	**SVR**	**Improvement (%)**
Cluster 0	New Cases	**0.128**	*0.182*	29.67%
	New Deaths	**0.129**	*0.147*	12.24%
Cluster 1	New Cases	**0.033**	*0.072*	54.16%
	New Deaths	**0.039**	*0.043*	9.3%
Cluster 2	New Cases	**0.034**	*0.199*	82.91%
	New Deaths	**0.036**	*0.109*	66.97%

The Improvement (%) was gathered by counting equation (15) on table 7. The Improvement (%) average from table 7 or SVR is 42.55 %.

4. CONCLUSION

LSTM model developed with the help of data clustering in this paper have better performance than the hyperparameter tuned SVR model with 29.11% improvement in RMSE and 16.24% improvement in MAE and the LSTM developed by Ayyoubzadeh, *et al.* [4] with 20.22% improvement in RMSE and 42.55% in MAE. The LSTM model also improved in terms of overfitting compared to Ayyoubzadeh, *et al.* [4] by 5 of 6 of our LSTM have lower distance between training and validation loss. The clustering-based training really helps our LSTM to perform better, therefore the proposed model give better prediction of COVID-19 in Indonesia especially in New Cases and New Deaths.

REFERENCES:

[1] K. C. Santosh, "COVID-19 prediction models and unexploited data," *J. Med. Syst.*, vol. 44, no. 9, pp. 1–4, 2020.

[2] S. Olivia, J. Gibson, and R. an Nasrudin, "Indonesia in the Time of Covid-19," *Bull. Indones. Econ. Stud.*, vol. 56, no. 2, pp. 143–174, 2020.

[3] L. J. Muhammad, M. M. Islam, S. S. Usman, and S. I. Ayon, "Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery," *SN Comput. Sci.*, vol. 1, no. 4, pp. 1–7, 2020.

[4] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R. N. Kalhori, "Predicting COVID-19 incidence through analysis of google trends data in iran: data mining and deep learning pilot study," *JMIR public Heal. Surveill.*, vol. 6, no. 2, p. e18828, 2020.

[5] S. Ballı, "Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods," *Chaos, Solitons & Fractals*, vol. 142, p. 110512, 2021.

[6] K. Park, Y. Choi, W. J. Choi, H.-Y. Ryu, and H. Kim, "LSTM-based battery remaining useful life prediction with multi-channel charging profiles," *Ieee Access*, vol. 8, pp. 20786–20798, 2020.

[7] K. P. Sinaga and M.-S. Yang, "Unsupervised K-means clustering algorithm," *IEEE access*, vol. 8, pp. 80716–80727, 2020.

[8] T.-M. Song and J. Song, "Prediction of risk factors of cyberbullying-related words in Korea: Application of data mining using social big data," *Telemat. Informatics*, vol. 58, p. 101524, 2021.

[9] D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat, "The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data," *Qual. Quant.*, vol. 56, no. 3, pp. 1283–1291, 2022.

[10] T. Li, A. Rezaeipanah, and E. M. T. El Din, "An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement," *J. King Saud Univ. Inf. Sci.*, vol. 34, no. 6, pp. 3828–3842, 2022.

[11] Z. Zhang *et al.*, "Solar radiation intensity probabilistic forecasting based on K-means time series clustering and Gaussian process regression," *IEEE Access*, vol. 9, pp. 89079–89092, 2021.

[12] H. Teichgraeber and A. R. Brandt, "Clustering methods to find representative periods for the optimization of energy systems: An initial framework and comparison," *Appl. Energy*, vol. 239, pp. 1283–1293, 2019.

[13] D. Li, Y. Zhao, and Y. Li, "Time-series representation and clustering approaches for sharing bike usage mining," *IEEE Access*, vol. 7, pp. 177856–177863, 2019.

[14] S. V Mashtalir, M. I. Stolbovyi, and S. V Yakovlev, "Clustering video sequences by the method of harmonic k-means," *Cybern. Syst. Anal.*, vol. 55, pp. 200–206, 2019.

[15] S. A. Abbas, A. Aslam, A. U. Rehman, W. A. Abbasi, S. Arif, and S. Z. H. Kazmi, "K-means and k-medoids: Cluster analysis on birth data collected in city Muzaffarabad, Kashmir," *IEEE Access*, vol. 8, pp. 151847–151855, 2020.

[16] S. Khare, M. K. Gourisaria, G. M. Harshvardhan, S. Joardar, and V. Singh, "Real Estate Cost Estimation Through Data Mining Techniques," in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1099, no. 1, p. 12053.

[17] G.-F. Fan, M. Yu, S.-Q. Dong, Y.-H. Yeh, and W.-C. Hong, "Forecasting short-term electricity load using hybrid support vector regression with grey catastrophe and random forest modeling," *Util. Policy*, vol. 73, p. 101294, 2021.

[18] N. J. Johannesen, M. Kolhe, and M. Goodwin, "Relative evaluation of regression tools for urban area electrical energy demand forecasting," *J. Clean. Prod.*, vol. 218, pp. 555–564, 2019.

[19] S. M. Hosseini, A. Saifoddin, R. Shirmohammadi, and A. Aslani, "Forecasting of CO2 emissions in Iran based on time series and regression analysis," *Energy Reports*, vol. 5, pp. 619–631, 2019.

[20] W. Xu, H. Peng, X. Zeng, F. Zhou, X. Tian, and X. Peng, "A hybrid modelling method for time series forecasting based on a linear regression model and deep learning," *Appl. Intell.*, vol. 49, pp. 3002–3015, 2019.

[21] Z. Zhang and W.-C. Hong, "Application of variational mode decomposition and chaotic grey wolf optimizer with support vector regression for forecasting electric loads," *Knowledge-Based Syst.*, vol. 228, p. 107297, 2021.

[22] S. Maldonado, A. Gonzalez, and S. Crone, "Automatic time series analysis for electric load forecasting via support vector regression," *Appl. Soft Comput.*, vol. 83, p. 105616, 2019.

[23] X. Xing, Y. Lin, H. Gao, and Y. Lu, "Wireless traffic prediction with series fluctuation pattern clustering," in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2021, pp. 1–6.

[24] M. Bilgili, N. Arslan, A. ŞEKERTEKİN, and A. YAŞAR, "Application of long short-term memory (LSTM) neural network based on deeplearning for electricity energy consumption forecasting," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 30, no. 1, pp. 140–157, 2022.

[25] E. Zhang, H. Li, Y. Huang, S. Hong, L. Zhao, and C. Ji, "Practical multi-party private collaborative k-means clustering," *Neurocomputing*, vol. 467, pp. 256–265, 2022.

[26] D. Amelia, T. N. Padilah, and A. Jamaludin, "Optimasi Algoritma K-Means Menggunakan Metode Elbow dalam Pengelompokan Penyakit Demam Berdarah Dengue (DBD) di Jawa Barat," *J. Ilm. Wahana Pendidik.*, vol. 8, no. 11, pp. 207–215, 2022.

[27] M. C. Yesilli, F. A. Khasawneh, and A. Otto, "Chatter detection in turning using machine learning and similarity measures of time series via dynamic time warping," *J. Manuf. Process.*, vol. 77, pp. 190–206, 2022.

[28] T. Chen, X. Shi, and Y. D. Wong, "A lane-changing risk profile analysis method based on time-series clustering," *Phys. A Stat. Mech. its Appl.*, vol. 565, p. 125567, 2021.

[29] M. Ratchagit and H. Xu, "A Two-Delay Combination Model for Stock Price Prediction," *Mathematics*, vol. 10, no. 19, p. 3447, 2022.

[30] Q. Li, D. Li, K. Zhao, L. Wang, and K. Wang, "State of health estimation of lithium-ion battery based on improved ant lion optimization and support vector regression," *J. Energy Storage*, vol. 50, p. 104215, 2022.

[31] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geosci. Model Dev.*, vol. 15, no. 14, pp. 5481–5487, 2022.

[32] A. Z. E. Kassem, "Does temperature affect COVID-19 transmission?," *Front. public Heal.*, vol. 8, p. 554964, 2020.