Optimizing the long short-term memory algorithm to improve the accuracy of infectious diseases prediction

Eko Sediyono[1], Sri Ngudi Wahyuni[2], Irwan Sembiring[1]
[1]Faculty of Information Technology, Satya Wacana Christian University, Salatiga, Indonesia
[2]Department of Informatics Management, Faculty of Computer Science, University of Amikom Yogyakarta, Yogyakarta, Indonesia

Article Info

Article history:

Received Jan 20, 2024
Revised Feb 26, 2024
Accepted Mar 5, 2024

Keywords:

Accuracy
Improvement
Infectious diseases
Optimized long short-term memory
Prediction

ABSTRACT

This study discusses the implementation of the proposed optimized long short-term memory (LSTM) to predict the number of infectious disease cases that spread in Central Java, Indonesia. The proposed model is developed by optimizing the output layer, which affects the output value of the cell state. This study used cases of four infectious diseases in Indonesia's Central Java Province, namely COVID-19, dengue, diarrhea, and hepatitis A. This model was compared to basic LSTM and MinMax schaler LSTM improvement to see the difference in the accuracy of each disease. The results showed a significant difference in the average prediction results with real cases between the three models. The main objectives of this study were: modifying the LSTM algorithm to predict the number of infectious disease cases to get a smaller residual value, comparing the results of the optimization accuracy of the LSTM algorithm with the LSTM algorithm in previous studies, and evaluating the use of spatial variables in applying infectious disease prediction models using the LSTM algorithm. The results found that the performance difference between the proposed optimization algorithm and the model in the previous study was obtained. The proposed LSTM optimization algorithm had an accuracy improvement of about 2% over the previous model.

Corresponding Author:

Eko Sediyono
Faculty of Information Technology, Satya Wacana Christian University
Notohamidjojo Street, Blotongan, Salatiga, Indonesia
Email: eko@uksw.edu

1. INTRODUCTION

Infectious diseases can pass from one individual to another, both in humans and animals. Infectious diseases are caused by biological agents such as pathogenic microorganisms (viruses, bacteria, and fungi) and parasites [1]. Over the past five years, infectious diseases have plagued several countries, including Indonesia. Tropical countries are more susceptible to certain contagious diseases due to several factors, such as climate, environmental conditions, and socioeconomic factors [2]. The tropics usually have warm and humid climates, creating favorable conditions for the survival and breeding of disease-carrying vectors such as mosquitoes and ticks. These vectors can transmit malaria, dengue, Zika, and chikungunya [3]. In addition, the tropics have biodiversity that can increase the risk of zoonotic diseases or diseases transmitted from animals to humans, such as the ebola virus and HIV. In addition to the above factors, many other factors cause tropical countries to be vulnerable to infectious diseases. Central Java is one of the provinces in Indonesia with a population of 36.52 million people and an area of 32.800.69 km². It has a population density of 1.135 people/square km, which triggers the massive spread of disease in this area [4]. In 2020, several infectious diseases became health problems in Central Java Province, including COVID-19, dengue, hepatitis, and diarrhea. Although there are few casualties, contagious diseases must be handled carefully and appropriately to optimally manage health services and facilities. It is necessary to predict infectious diseases in preparation for community health facilities so that there are no death cases from infectious diseases [5]. The main problem in infectious disease prediction is the gap between the prediction results and what occurs in the field; therefore, it cannot be used for decision-making by related parties. Frequent changes in engine settings in the forecast model cause prediction results to change. The accuracy of the selection and approach of the forecasting model dramatically affects the prediction results. This study explains the optimization of long short-term memory (LSTM) algorithms for infectious disease prediction, especially time series data [6].

This study investigated the effect of modifying the LSTM algorithm on the output layer to predict the number of infectious disease cases using numerical and spatial variables in prediction parameters. The algorithm was modified by reducing the state cell output value at the LSTM algorithm's output gate. Optimization of the output layer in the LSTM model was aimed at reducing the resetting of the machine during the learning model process. The resetting process burdened machine performance and parsed the deviation of prediction results [7]. Based on this problem, the main objectives of this study were: i) modifying the LSTM algorithm to predict the number of infectious disease cases to get a smaller residual value, ii) comparing the results of the optimization accuracy of the LSTM algorithm with the LSTM algorithm in previous studies, and iii) evaluating the use of spatial variables in applying infectious disease prediction models using the LSTM algorithm.

This research focused on how modifying the LSTM algorithm results in high accuracy in predicting infectious diseases. This predictive model helped the government and others take strategic steps to control infectious diseases in Indonesia [8], especially in Central Java Province, such as providing vaccines and hospital services, mapping health workers, and providing health budgets [9]. Applying this optimation predictive model helps the government take early action to prevent infectious diseases that occur periodically [10]. The second part of the study investigated some modifications of existing LSTM algorithms, specifically on infectious disease prediction. The third part describes the process of LSTM algorithm modification to increase the accuracy value of prediction models, the last part tests the model performance by comparing accuracy results, and part five summarizes a series of works done on this study.

2. LITERATURE REVIEW

2.1. Infectious diseases prediction model approach

The main problem in infectious disease prediction is the inaccuracy of prediction results close to actual cases, so they cannot be used in decision-making by interested parties. Frequent changes in setting the machine in the forecasting model cause the prediction results to change. The accuracy of the selection and approach of the forecasting model dramatically influences the prediction results. Several approaches to time series-data prediction models for infectious diseases have been widely studied, including statistical approaches such as autoregressive integrated moving average (ARIMA) [11], [12], exponential smoothing (ES) [13], vector autoregression (VAR) [14], generalized autoregressive conditional heteroskedasticity (CARCH) [15], seasonal ARIMA [16], and prophet algorithm [17], [18]. However, statistical approaches perform poorly for distance and small amounts of data. In contrast, adding infectious disease data is erratic, sometimes adding a lot of data or vice versa. Next is the mathematical approach, among others, the susceptible-infectious-recovered (SIR) [19], [20], susceptible-exposed-infectious-recovered (SEIR) [21], [22]. Nevertheless, the accuracy of this model's prediction results depends on the data quality and the amount of data and requires many prediction parameters.

There are several machine learning approaches, including logistic regression [23], random forest [24], support vector machine (SVM) [25], and naïve Bayes [26]. However, this model cannot present trend predictions for time series data. Therefore, the model performance decreases. Looking at model performance decreases, deep learning is designed as a viable model for time series data prediction because it can process temporal and time series data [27].

There are several prediction models in deep learning, one of which is LSTM-neural networks [28]. The LSTM model itself has several variants, including vanilla LSTM, stacked LSTM [9], bidirectional LSTM [29], encoder-decoder LSTM [30], recurrent neural networks (RNN) [31], LSTM to be able to present trend predictions well [32]. LSTM is one of the most influential and widely used models for time series forecasting. LSTM is a repetitive neural network designed to capture and model sequential dependencies in time series data whose patterns may be scattered over time. The number of cases in infectious diseases is time series data that changes all the time, so it is appropriate to use LSTM models.

Nonetheless, the data patterns and the amount of data are also essential when applying LSTM models for the prediction [33]. As a solution, this research conducts LSTM optimization to solve these problems. The proposed optimized long short-term memory (popLSTM) optimizes the output layer (*ot*) by placing tanh as a subtraction of the value of 1 on that layer so that the hidden state (*ht*) in the output layer can suppress the model error value.

2.2. Long short-term memory prediction model applied in recent years

The LSTM model is widely used in predicting infectious disease time-series data, including by Yang *et al.* [34] predicting cases of tuberculosis. The study used maximum temperature, average relative humidity, local financial budgets, monthly sunshine percentages, and sunshine hours as predictive variables. The experimental results showed that LSTM performed better than other models. LSTM reduced the model's error in predicting by 13-16%. In this study, data patterns greatly affected the model's accuracy, so we had to pay attention to data patterns and the data used to train the model [35]. Furthermore, Gu *et al.* [6] used LSTM to predict palm and foot disease in animals in China. This study used flexibility, temperature, air pressure, and wind speed variables in making predictions. Results showed that LSTM could predict the upcoming incidence of foot-and-mouth disease in Guangxi, China. Leveraging these variables enabled LSTM models to grasp intricate relationships among multiple variables quickly and outperformed models relying solely on a single variable [36].

Chae *et al.* [37] conducted a study on the prediction of the spread of Chickenpox disease in Korea using LSTM; the prediction results were very accurate compared to the ARIMA model. The prediction results helped eliminate reporting delays in the monitoring system for the spread of infectious diseases in the community, thereby minimizing health costs. Research has also shown that weather variables, internet big data, and deep learning could help predict infectious diseases more effectively [38]. Guo *et al.*[39] predicted acute hepatitis E in Shandong, China, using LSTM, SVM, and ARIMA models. The results stated that LSTM had the highest accuracy value compared to the other two models. The study predicted hepatitis E based on machine learning models using three models. They were ARIMA, SVM, and LSTM-RNN. Experimental data were obtained from the monthly incidence and number of hepatitis E cases from January 2005 to December 2017, with selected data from July 2015 to December 2017. Three metrics were applied to compare model performance: root mean square error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE). The results showed that ARIMA and SVM predicted a monthly incidence of hepatitis E cases. However, LSTM was superior to both because LSTM can read all modeling and data patterns [40].

Wang *et al.* [41] implemented LSTM to predict the number of malaria cases in Yunnan, China. The results showed that the LSTM model had better accuracy than other models. When using LSTM in this research, the effectiveness and quantity of the data played a significant role. Hence, it was crucial to consider the data preprocessing and feature engineering phases when employing LSTM models for predictive purposes [36]. Xu *et al.* [42] predicted a dengue outbreak one month ahead in China using the LSTM model. The prediction parameters used are demographics, the number of new cases, deaths, hospitalized cases, and the climate in China. The results improved the generalizability of the model in the region, and the model could estimate the number of dengue cases more precisely. Therefore, the model is recommended to be used as a forecasting model for similar infectious diseases. Based on the description above, the LSTM model can predict infectious diseases well. However, the study still used small datasets and has yet to consider large numbers of datasets [43].

2.3. Optimization long short-term memory method research

In some research on LSTM optimization on prediction, Saleh *et al*. [44] conducted predictions using optimized LSTM. In this study, prediction accuracy improvement was carried out by combining LSTM with Bayesian. The experimental results showed that this model had a very high accuracy value compared to models that do not use Bayesian. However, this model had the disadvantage that it has yet to be tested using data other than previously tested datasets. Furthermore, it took hard work for machines to train data using this model. Thus, further studies are needed to prove the accuracy value of the optimized model. Yan [45] predicted the number of COVID-19 cases in several countries using LSTM optimization in the MinMax scaler data process. In this experiment, the accuracy value was better than the accuracy of the basic LSTM model and logistic regression using a deep learning approach. However, the optimization must be manually installed in this model because it cannot be placed at the entire data distance. Hence, it is not possible in terms of automation at a particular data distance.

Ewees *et al.* [46] optimized the LSTM algorithm by combining heap-based optimizer (HBO) using LSTM as an in-depth learning technique to predict wind energy derived from multiple wind turbines. Metaheuristic optimization algorithms, like the HBO, were applied to train the LSTM and improve the prediction's performance. This study evaluated HBO–LSTM, which was developed on four data groups of La Haute Borne in France, and compared them with particle swarm optimization-LSTM, differential evolution-LSTM, genetic algorithm-LSTM, salp swarm algorithm-LSTM, sine cosine algorithm-LSTM, and grey wolf optimization-LSTM. The result indicated that there was a significant increase in HBO's performance.

Zhou *et al.* [47] improved the accuracy of LSTM models by optimizing deep-learning approaches. In this experiment, COVID-19 data from WHO was used to compare the prediction accuracy of COVID-19 cases in 12 countries worldwide. This experiment compared the results of a data mining approach with a deep learning approach to predict infectious diseases. The models compared were LSTM and Bi-LSTM. The results showed that predictive experiments using a deep learning approach had better accuracy and could handle complex things than data mining. However, this approach also had weaknesses because not all data on infectious diseases can be predicted using this approach model. Therefore, further studies are needed to prove it.

3. MATERIAL AND METHOD

The experimental steps in this study are presented in Figure 1. The first step was infectious disease data collection in Central Java Province, Indonesia, consisting of four infectious diseases: COVID-19, dengue, diarrhea, and hepatitis A. Data was taken from the Hospital Laboratory of Dr. Kariadi Semarang, Indonesia, and the Central Bureau of Statistics Central Java Province, Indonesia. The second step was data preprocessing. It was conducted by dividing data into training and test data sets. We divided the training and test data by 80% and 20%. The input in this training was time series data, with prediction variables, the number of confirmed cases, the number of recovered patients, the number of patients who died, population density, latitude, and longitude. The model output was time series data, the number of confirmed cases. The model was intended to see the prediction of the number of patients with the four diseases in the next seven days.

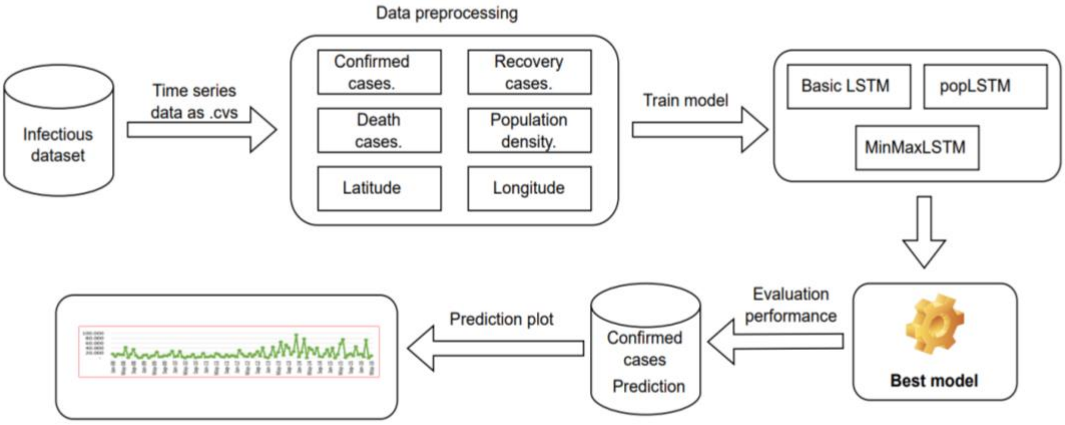

Figure 1. Proposed framework

The third step was the data training process. It preprocessed infectious disease datasets using three models: basic LSTM, MinMaxLSTM, and popLSTM. We conducted data training for 7 days, and predictions were carried out for the next 7 days, December 25-31, 2022. Data training was carried out for 7 days because the incubation period of infectious diseases occurred on the seventh day. The tools used in this experiment are Google Colaboratory and several libraries from Scikitlearn and Tensorflow to see prediction results with actual data. The fourth step was predicting confirmed cases of infectious diseases using basic LSTM, MinMaxLSTM, and popLSTM models. At this step, data training was carried out using three models, basic LSTM, MinMaxLSTM, and popLSTM, to see the accuracy of the proposed model. These models have been widely used in previous studies so that the experimental results can detect the residual values of the three models and determine the model with the best accuracy value.

The fifth step was performing accuracy testing using MAE and R square test parameters. In this step, the test results were analyzed, and the lowest residual values of the three models were compared. The lowest residual value indicated the best model in the experiment. The sixth step was plot prediction. The plot in this experiment illustrated three things: the prediction results of the three models, the difference in model accuracy, and the difference in model residual values seen from the results of MAE and R square. Furthermore, the plot results were analyzed to conclude the recommended model for infectious disease prediction in Central Java.

3.1. Long short-term memory

LSTM is an artificial neural network architecture that handles sequential data, such as speech, text, and time series data. LSTM was introduced in 1997 by Hochreiter and Schmidhuber [48] and has since become a popular and powerful tool for various applications, including natural language processing, speech recognition, and image captions. LSTM is one of the RNN algorithms that can process data sequences by maintaining an internal state and being able to capture long-term dependencies. Memory cells in LSTM make it possible to recall or forget information selectively over long periods. Memory cells are controlled by three gates: input, output, and forget. These three gates regulate the flow of information in and out of the cell. This gate allows LSTM to selectively store or discard information based on the relevance of the task. The three gates are presented in (1) to (6):

$$ft = \sigma(W_f.[h_{t-1}, x_t] + b_f \tag{1}$$

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i \tag{2}$$

$$\tilde{C_t} = tanh\ (W_c.[h_{t-1}, x_t] + b_c \tag{3}$$

$$C_t = f_t * c_{t-1} + i_t * \tilde{C}_t \tag{4}$$

$$ot = \sigma(w_o.[h_{t-1}, x_t] + b_o \tag{5}$$

$$h_t = o_t\ {*tanh}\ tanh\ (C_t) \tag{6}$$

ft is the forget gate determining which information to pay attention to and which could be ignored. Meanwhile, ⬚$_{it}$ shows the output gate at t, and σ represents the sigmoid function. W_f represents the weight value for the forget gate. The $h_{(t-1)}$ is the hidden state, and $b_f$represents the forgot gate bias value [49]. i_t represents the input gate that updates the cell status, and $\tilde{C_i}$ indicates the cell state operation. Figure 2 is the architecture of LSTM.

3.3. Proposed optimized long short-term memory method

This study proposes LSTM optimization by degrading the value of the cell state (*ht*). This value is obtained by lowering the *ot* value at the output gate to <0.5 so that this value will trigger the cell state value to be low. The popLSTM is presented in Figure 3. Figure 3 shows popLSTM (7) to (12). Were:

$$1 - ot = 1 - \ \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{7}$$

In which:

$$\sigma(x) = \frac{1}{1+\epsilon^{-x}} \tag{8}$$

Accordingly:

$$1 - ot = \frac{\epsilon^{-Wo[h_{t-1},x_t]}\cdot\epsilon^{-bo}}{1+\epsilon^{-Wo[h_{t-1},x_t]}.\epsilon^{-bo}} \tag{9}$$

The result shows that. ⬚$_{It}$ is:

$$ot = 1 - \frac{\epsilon^{-Wo[h_{t-1},x_t]}\cdot\epsilon^{-bo}}{1+\epsilon^{-Wo[h_{t-1},x_t]}.\epsilon^{-bo}} \tag{10}$$

$$ot = 1 - \frac{\epsilon^{-Wo[h_{t-1},x_t]}\cdot\epsilon^{-bo} - \epsilon^{-Wo[h_{t-1},x_t]}\cdot\epsilon^{-bo}}{1+\epsilon^{-Wo[h_{t-1},x_t]}.\epsilon^{-bo}} \tag{11}$$

$$ot = \frac{1}{1+\epsilon^{-Wo[h_{t-1},x_t]}.\epsilon^{-bo}} \tag{12}$$

The final predicted value of the LSTM has shown dependency on the final value of h_t. Meanwhile, the last h_t Value depends on the previous *ot* value. In this research, when the value of *ot*<0.5, the h_t value decreases, subtracting the value of 1 with *ot* or $1 - \sigma(W_o.[h_{t-1}, x_t] + b_o)$. Therefore, the last *ot* to value decreases in number, and the h_t value will automatically do the same due to multiplying with tanh on C_t.

Value 1 of the above function is the parameter that was added to reduce the value of h_1 to make a lower output *o*f of going to h_1. The result will range from 0 to 1, and the output h_1 will also have a lower value. When the value of information gate o_t is less than 0.5, gate o_t will tend to cut off or inhibit most of the incoming information from h_t. A value less than 0.5 leads to special treatment in the flow of information through the LSTM cell at a specific time and result in output restriction on h_t.

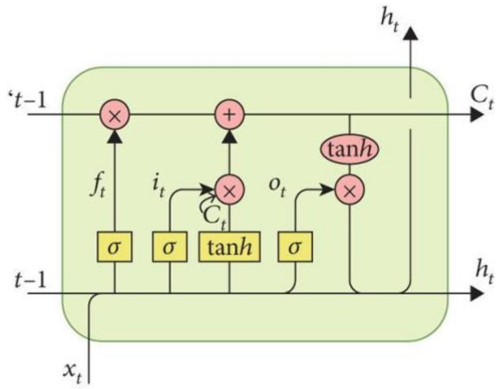

Figure 2. The architecture of LSTM

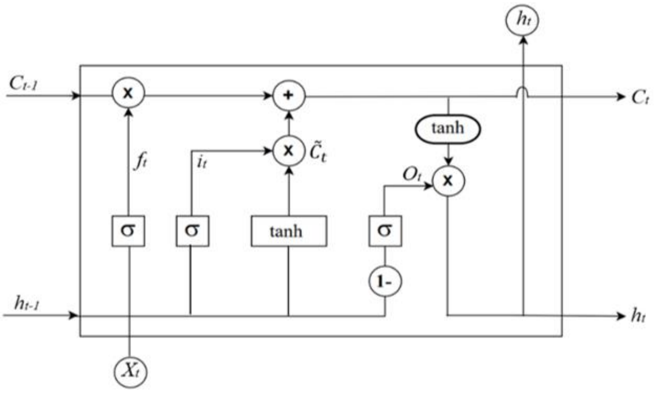

Figure 3. The popLSTM architecture

3.4. Evaluation performance

3.4.1. R square

In this study, R square testing is used to get an overview of model validation where if the value is close to 1, then the model is considered close to the data. The R square function is presented in (13) to (15).

$$R^2 = \frac{SSregression}{SStotal} \tag{13}$$

Where *SSregression* is the coefficient of determination between 0 and 1. Mathematically, it is calculated as the sum of the squared differences between the predicted and mean values of the dependent variable, and *SStotal* represents the total sum squared, representing the sum of squares of the difference between each observation value of the dependent variable and the average of the dependent variable [50].

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{14}$$

$\hat{y}$ is the prediction value, the average value of the dependent variable.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{15}$$

where y_i is the dependent variable and $\bar{y}$ is the average value.

3.4.2. Mean absolute error

The MAE test measures the average of the absolute values of the difference between the predicted and actual values. The value of the MAE matrix ranges from 0 to infinity. Next, MAE is calculated based on the absolute difference between predicted and actual values [50]. The absolute diffference is calculated by taking the difference between the prediction and the actual value. The MAE equation is presented in (16), where y_j is the prediction value, $y\,\hat{j}$ is the real value, and n is the sum of data.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \tag{16}$$

Where y_j is the prediction value, $y\,\hat{j}$ is the real value, and n is the sum of data. The lower the MAE value, the better the model makes accurate predictions.

4. RESULT

This section discusses the steps to implement popLSTM results based on prediction results and accuracy test results. Furthermore, the accuracy test results were compared to the experimental results from the other two models. The comparison results will be used as recommendations to interest parties in mapping health mechanisms to prevent infectious diseases in Central Java Province, Indonesia. This contagious disease is chosen because there is climate change in Central Java, which triggers the disease to occur at any time, especially from the transition period of the dry season to the rainy season.

4.1. Data selection and processing

The dataset in this study used public data sourced from Laboratories of Dr. Kariadi Hospital in Semarang, Central Java, Indonesia, and the Central Bureau of Statistics Central Java Province. Data was taken five years earlier, from 2019 to 2022. The infectious disease data that became sample data in the data training process were COVID-19, dengue, diarrhea, and hepatitis A. These four infectious diseases have plagued Central Java Province for five years. The prediction parameters used are the number of confirmed cases, the number of recovered patients, the number of patients who died, population density, latitude, and longitude. Spatial parameters such as population density, latitude, and longitude were selected to see patterns of disease distribution by region. The prediction result was the number of new cases in the next seven days. The population density, latitude, and longitude data was obtained publicly at [5].

To improve machine performance, we placed the modified model on the output layer, or *ot*, to filter information from cell state or *Ct*. This modification filtered information from the previous gate to a value of <0.5 so that the information produced by *ht* is efficient. Reducing the output of information by half of what it was before makes the value of information more accurate.

4.2. Parameter

In this experiment, we set a time series of 7 days for data training and 7 days for predicting COVID-19, dengue hemorrhagic fever (DHF), diarrhea, and hepatitis A for infectious diseases in Central Java. Predictions were carried out using a trained model that has been modified to get a prediction of the number of confirmed cases of the four infectious diseases. The placement of modified models aimed to reduce machine performance until the training and learning processes also decreased. This process resulted in a decrease in residual value and an increase in accuracy value.

4.3. Prediction result

We analyzed and compared the performance of the three predictive models by looking at the number of infectious disease cases. The predicted results are illustrated in Table 1. Next, we compared the performance of the three models to see which one had the best performance. The results of the model performance comparison are presented in Table 2. The comparison results showed that there was an increase in the proposed optimization algorithm. The popLSTM algorithm has an accuracy improvement of about 2% over the previous model.

Table 1. The Central Java infectious diseases confirmed cases prediction comparison

Date	Diseases	Real cases	Basic LSTM	MinMaxLSTM	PopLSTM
25/12/2022	Covid-19	655117	655133	655070	655117
26/12/2022		655201	655130	655120	655201
27/12/2022		655248	655158	655130	655248
28/12/2022		655283	655249	655199	655283
29/12/2022		655251	655275	655257	655251
30/12/2022		655338	655336	655321	655338
31/12/2022		655409	655383	655323	655409
25/12/2022	Dengue	165	212	179	179
26/12/2022		177	216	203	190
27/12/2022		122	166	153	143
28/12/2022		124	161	134	142
29/12/2022		133	173	153	147
30/12/2022		170	182	193	183
31/12/2022		177	209	196	195
25/12/2022	Diarrhea	2906	2941	2937	2935
26/12/2022		2957	3012	2989	2982
27/12/2022		2804	2831	2826	2821
28/12/2022		2677	2705	2703	2697
29/12/2022		2906	2959	2930	2926
30/12/2022		3059	3075	3081	3088
31/12/2022		2651	2693	2694	2677
25/12/2022	Hepatitis A	6	26	17	16
26/12/2022		7	13	22	17
27/12/2022		4	22	14	13
28/12/2022		6	16	15	15
29/12/2022		6	15	18	16
30/12/2022		4	23	15	11
31/12/2022		6	27	16	11

Table 2. Evaluation model comparison

Diseases	Model	MAE	R^2
COVID-19	Basic LSTM	47.73	0.993
	MinMaxLSTM	24.97	0.704
	popLSTM	22.55	0.993
Dengue	Basic LSTM	14.26	0.922
	MinMaxLSTM	40.92	0.995
	popLSTM	16.53	0.877
Diarrhea	Basic LSTM	14.95	0.997
	MinMaxLSTM	9.57	0.951
	popLSTM	31.49	0.997
Hepatitis A	Basic LSTM	12.31	0.934
	MinMaxLSTM	11.16	0.998
	popLSTM	7.01	0.975

Table 1 shows the predicted results of three models for the next seven days. The performance of the model increased by 2% as compared to the other algorithms as shown in Table 1. There were also differences in the outcome average of each model. The average deviation of COVID-19 disease prediction results for the basic LSTM, MinMaxLSTM, and popLSTM models in the next seven days is 0.17%, 0.13%, and 0.007%. Therefore, popLSTM has a better-predicted deviation value than other models; for dengue disease, the predicted residue results are 24%, 11%, and 9%, so popLSTM has the smallest deviation value compared to other models. The predicted residue results for diarrhea are 1.28%, 1.002%, and 0.832%, so popLSTM has the smallest residue value compared to other models. Furthermore, the residue values of hepatitis A disease are 26.4.4%, 20%, and 15%.

Based on the experiments in this study, as shown in Table 1, we found that modifications to the o_t layer and the placement of spatial variables affected the model's residual and accuracy values. It can be seen that there was a residual value difference of between 5-6% at a small data distance, while for a significant data distance, we found a residual difference of up to 2%. This modified model was better than the previous LSTM-modified prediction model [45]. The popLSTM model can be recommended as an infectious disease prediction model.

4.4. Evaluation result

This evaluation stage tests the model to see its accuracy and compares it to see which model has the best accuracy. It is recommended for infectious disease time series-data forecasting. The results of model testing are presented in Table 2. Table 2 compares model evaluation results using MAE and R square. Based on Table 2, the popLSTM model has a higher accuracy value than other models. Based on the accuracy testing results in Table 2, the average MAE value for the Basic LSTM model is 22.3125, for the MinMax LSTM model 21.655, and for the popLSTM model is 19.395, so there is an increase in accuracy of 2.29% in the popLSTM model than MinMax LSTM model. See the average evaluation results of the four infectious diseases presented in Table 3.

Table 3. Average model evaluation results

Model	Average of MAE	Average of R Square
Basic LSTM [5]	22.31	0.9030
MinMaxLSTM [44]	21.65	0.9550
popLSTM	19.395	0.9760

Table 3 shows the model evaluation's average result using MAE and R Square. It can be seen that popLSTM has an accuracy value of 97.6%, the MinMaxLSTM model is 95.5%, and the basic LSTM model has an accuracy value of 90% for the prediction of four infectious diseases, especially time series data as shown in Table 3. PopLSTM has superior accuracy values compared to other models. While the average MAE value of popLSTM is 14.44, the average value of M.A.E. MinMaxLSTM is 19.07, and the average value of MAE in basic LSTM is 25.5, the error value in the popLSTM model is lower than the error value in other models. The comparison in Table 3 is presented in Figure 4.

Table 3 and Figure 4 compared our best models with models from previous studies using infectious disease datasets. The comparison showed that the present model outperformed existing prediction models and exceeded current conditions using numerical and spatial variables. This study showed that modifying the output layer on LSTM provided better accuracy values than previous studies. Using spatial variables, especially longitude and latitude, provided significant results because the model could also support higher accuracy values. After all, it could study the geographical patterns of an area so that it could map disease-prone locations [51].

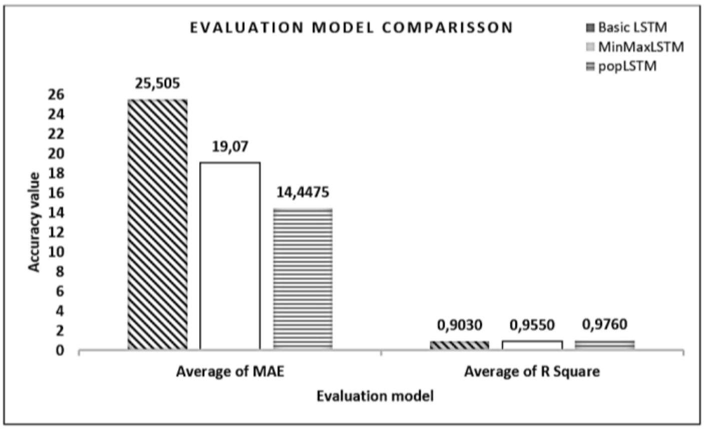

Figure 4. The comparison of average MAE and R square all models

The limitation of the study is that disease data is not publicly accessible, and data is incomplete in some government web sources. In addition, publicly accessible data is less updated. In future studies, spatial variables such as temperature, humidity, weather, and several other spatial variables can be added to obtain more detailed prediction values and disease range locations based on regional geography.

5. CONCLUSION

Based on observations of experimental results, this study concludes that the popLSTM model is suitable for predicting time series data. The results show that our proposed LSTM-modified model has an increase in accuracy value of more than 2%. This result was better than the previous research-modified model. The proposed modification of the LSTM model was trained for seven days by adjusting the incubation period of four infectious diseases: COVID-19, dengue, diarrhea, and hepatitis A. This proposed model made short-term predictions, which were seven days ahead. This value was seen by comparing the residual values of all models. We found that using spatial variables, population density, longitude, and latitude played an essential role in this prediction. The experimental results showed that the predicted value of the number of confirmed cases of infectious diseases using the proposed LSTM modification model was better than the LSTM modification models in previous studies. This value was indicated by the average MAE value for the basic LSTM model, which is 22.3125, for the MinMax LSTM model, 21.655, and the popLSTM model is 19.395, so there was an increase in accuracy of 2.29% in the popLSTM model than previous LSTM modification model. The limitation of the study was that disease data was not publicly accessible. Besides, data from some government web sources was incomplete. In addition, publicly accessible data was less updated. Thus, adjustments to the engine were needed during the prediction process. Future research should explore other time series prediction algorithms capable of performing automatic machine adjustments to reduce training time. Furthermore, it should explore optimizing other predictive models using more spatial variables to improve accuracy in disease prediction models for big and small data.

ACKNOWLEDGEMENTS

We thank to Satya Wacana Christian University, Salatiga, Indonesia, that have supported this research, by research grant number 095/SPK-PT/RIK/8/2023. It is the collaboration research between students and professors in the Faculty of Information Technology doctoral program.

REFERENCES

[1] L. S. Fischer, G. Mansergh, J. Lynch, and S. Santibanez, “Addressing disease-related stigma during infectious disease outbreaks,” *Disaster Medicine and Public Health Preparedness*, vol. 13, no. 5–6, pp. 989–994, Dec. 2019, doi: 10.1017/dmp.2018.157.

[2] S. Dash, C. Chakraborty, S. K. Giri, and S. K. Pani, “Intelligent computing on time-series data analysis and prediction of COVID-19 pandemics,” *Pattern Recognition Letters*, vol. 151, pp. 69–75, Nov. 2021, doi: 10.1016/j.patrec.2021.07.027.

[3] C. Edussuriya, S. Deegalla, and I. Gawarammana, “An accurate mathematical model predicting number of dengue cases in tropics,” *PLoS Neglected Tropical Diseases*, vol. 15, no. 11, Nov. 2021, doi: 10.1371/journal.pntd.0009756.

[4] S. N. Wahyuni, E. Sediyono, and I. Sembiring, “Indonesian covid-19 future forecasting based on machine learning approach,” *2021 3rd International Conference on Electronics Representation and Algorithm (ICERA)*, pp. 104–108, Jul. 2021, doi: 10.1109/ICERA53111.2021.9538672.

[5] S. N. Wahyuni, E. Sediono, I. Sembiring, and N. N. Khanom, “Comparative analysis of time series prediction model for forecasting

COVID-19 trend," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 1, pp. 595–605, Oct. 2022, doi: 10.11591/ijeecs.v28.i1.pp595-605.

[6] J. Gu *et al.*, "A method for hand-foot-mouth disease prediction using GeoDetector and LSTM model in Guangxi, China," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[7] X. H. Le, H. V. Ho, G. Lee, and S. Jung, "Application of long short-term memory (LSTM) neural network for flood forecasting," *Water*, vol. 11, no. 7, 2019, doi: 10.3390/w11071387.

[8] S. Rath, A. Tripathy, and A. R. Tripathy, "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 5, pp. 1467–1474, 2020, doi: 10.1016/j.dsx.2020.07.045.

[9] S. Hansun, V. Charles, and T. Gherman, "The role of the mass vaccination programme in combating the COVID-19 pandemic: an LSTM-based analysis of COVID-19 confirmed cases," *Heliyon*, vol. 9, no. 3, Mar. 2023, doi: 10.1016/j.heliyon.2023.e14397.

[10] L. Qin *et al.*, "Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index," *International Journal of Environmental Research and Public Health*, vol. 17, no. 7, 2020, doi: 10.3390/ijerph17072365.

[11] P. Somboonsak, "Forecasting dengue fever epidemics using ARIMA model," *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference*, pp. 144–150, 2019, doi: 10.1145/3375959.3375970.

[12] M. S. Peiris and B. J. C. Perera, "On prediction with fractionally differenced ARIMA models," *Journal of Time Series Analysis*, vol. 9, no. 3, pp. 215–220, 1988. doi: 10.1111/j.1467-9892.1988.tb00465.x.

[13] A. M. C. H. Attanayake, S. S. N. Perera, and U. P. Liyanage, "Exponential smoothing on forecasting dengue cases in Colombo, Sri Lanka," *Journal of Science*, vol. 11, no. 1, 2020, doi: 10.4038/jsc.v11i1.24.

[14] F. Khan, A. Saeed, and S. Ali, "Modelling and forecasting of new cases, deaths and recover cases of COVID-19 by using vector autoregressive model in Pakistan," *Chaos, Solitons and Fractals*, vol. 140, Nov. 2020, doi: 10.1016/j.chaos.2020.110189.

[15] E. Nkoro and A. K. Uko, "A generalized autoregressive conditional heteroskedasticity model of the impact of macroeconomic factors on stock returns: empirical evidence from the Nigerian stock market," *International Journal of Financial Research*, vol. 4, no. 4, 2013, doi: 10.5430/ijfr.v4n4p38.

[16] P. Somboonsak, "Time series analysis of dengue fever cases in Thailand utilizing the SARIMA model," *Proceedings of the 2019 7th International conference on information technology: IoT and smart city*, pp. 439–444, 2019, doi: 10.1145/3377170.3377215.

[17] 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, 2021, doi: 10.1016/J.PROCS.2021.01.036.

[18] A. K. Gupta, V. Singh, P. Mathur, and C. M. T. -Gonzalez, "Prediction of COVID-19 pandemic measuring criteria using support vector machine, prophet and linear regression models in Indian scenario," *Journal of Interdisciplinary Mathematics*, vol. 24, no. 1, 2021, doi: 10.1080/09720502.2020.1833458.

[19] I. Rahimi, A. H. Gandomi, P. G. Asteris, and F. Chen, "Analysis and prediction of covid-19 using sir, seiqr and machine learning models: Australia, italy and uk cases," *Information*, vol. 12, no. 3, pp. 1–25, Mar. 2021, doi: 10.3390/info12030109.

[20] B. Malavika, S. Marimuthu, M. Joy, A. Nadaraj, E. S. Asirvatham, and L. Jeyaseelan, "Forecasting COVID-19 epidemic in India and high incidence states using SIR and logistic growth models," *Clinical Epidemiology and Global Health,* vol. 9, 2020, pp. 26–33, 2021, doi: 10.1016/j.cegh.2020.06.006.

[21] P. Bedi, S. Dhiman, P. Gole, N. Gupta, and V. Jindal, "Prediction of COVID-19 trend in India and its four worst-affected states using modified SEIRD and LSTM models," *SN Computer Science*, vol. 2, no. 3, May 2021, doi: 10.1007/s42979-021-00598-5.

[22] M. Y. Li and J. S. Muldowney, "Global stability for the SEIR model in epidemiology," *Mathematical Biosciences,* vol. 125, no. 2, pp. 155-164, 1995.

[23] F. W. Wibowo, "Prediction modelling of covid-19 outbreak in indonesia using a logistic regression model," *Journal of Physics: Conference Series,* vol. 1803, no. 1, 2019, doi: 10.1088/1742-6596/1803/1/012015.

[24] E. Mussumeci, "A machine learning approach to dengue forecasting: comparing LSTM, random forest and Lasso," P.hD. Thesis, Escola de Matemática Aplicada, Brazil, 2018. [Online]. Available: https://bibliotecadigital.fgv.br/dspace/handle/10438/24093

[25] A. Khakharia *et al.*, "Outbreak prediction of covid-19 for dense and populated countries using machine learning," *Annals of Data Science*, vol. 8, 2020, doi: 10.1007/s40745-020-00314-9.

[26] A. Al-Hashedi *et al.*, "Ensemble classifiers for arabic sentiment analysis of social network (twitter data) towards covid-19-related conspiracy theories," *Applied Computational Intelligence and Soft Computing*, vol. 2022, no. 1, 2022, doi: 10.1155/2022/6614730.

[27] I. E. Livieris, S. Stavroyiannis, E. Pintelas, and P. Pintelas, "A novel validation framework to enhance deep learning models in time-series forecasting," *Neural Computing and Applications,*, vol. 32, no. 23, pp. 17149–17167, Dec. 2020, doi: 10.1007/s00521-020-05169-y.

[28] R. K. Pathan, M. Biswas, and M. U. Khandaker, "Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model," *Chaos, Solitons & Fractals*, vol. 138, 2020. doi: 10.1016/j.chaos.2020.110018.

[29] F. Li, G. Ma, S. Chen, and W. Huang, "An ensemble modeling approach to forecast daily reservoir inflow using bidirectional long- and short-term memory (Bi-LSTM), variational mode decomposition (VMD), and energy entropy method," *Water Resources Management,* vol. 35, no. 9, pp. 2941–2963, Jul. 2021, doi: 10.1007/s11269-021-02879-3.

[30] S. Verma and R. K. Gazara, "Big data analytics for understanding and fighting COVID-19," *Studies in Computational Intelligence*. Springer Singapore, pp. 333–348, 2020. doi: 10.1007/978-981-15-8534-0_17.

[31] S. Chae, S. Kwon, and D. Lee, "Predicting infectious disease using deep learning and big data," *International Journal of Environmental Research and Public Health*, vol. 15, no. 8, Aug. 2018, doi: 10.3390/ijerph15081596.

[32] X. Song *et al.*, "Time-series well performance prediction based on long short-term memory (LSTM) neural network model," *Journal of Petroleum Science and Engineering*, vol. 186, Mar. 2020, doi: 10.1016/j.petrol.2019.106682.

[33] A. M. Rather, "LSTM-based deep learning model for stock prediction and predictive optimization model," *EURO Journal on Decision Processes*, vol. 9, 2021, doi: 10.1016/j.ejdp.2021.100001.

[34] E. Yang, H. Zhang, X. Guo, Z. Zang, Z. Liu, and Y. Liu, "A multivariate multi-step LSTM forecasting model for tuberculosis incidence with model explanation in Liaoning Province, China," *BMC Infectious Diseases*, vol. 22, no. 1, pp. 1–13, 2022.

[35] A. Lawi, H. Mesra, and S. Amir, "Implementation of long short-term memory and gated recurrent units on grouped time-series data to predict stock prices accurately," *Journal of Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00597-0.

[36] H. Ji, Y. Lou, S. Cheng, Z. Xie, and L. Zhu, "An advanced long short-term memory (LSTM) neural network method for predicting rate of penetration (ROP)," *ACS Omega*, vol. 8, no. 1, pp. 934-945, 2022, doi: 10.1021/acsomega.2c06308.

[37] S. Chae, S. Kwon, and D. Lee, "Predicting infectious disease using deep learning and big data," *International Journal of Environmental Research and Public Health*, vol. 15, no. 8, 2018.

[38] M. Hawas, "Generated time-series prediction data of COVID-19′s daily infections in Brazil by using recurrent neural networks," *Data in Brief*, vol. 32, 2020. doi: 10.1016/j.dib.2020.106175.

[39] Y. Guo, Y. Feng, F. Qu, L. Zhang, B. Yan, and J. Lv, "Prediction of hepatitis E using machine learning models," *PLoS One*, vol. 15, no. 9, Sep. 2020, doi: 10.1371/journal.pone.0237750.

[40] Y. Gautam, "Transfer learning for COVID-19 cases and deaths forecast using LSTM network," *ISA Transactions*, vol. 124, pp. 41–56, May 2022, doi: 10.1016/j.isatra.2020.12.057.

[41] M. Wang *et al.*, "A novel model for malaria prediction based on ensemble algorithms," *PLoS One*, vol. 14, no. 12, 2019, doi: 10.1371/journal.pone.0226910.

[42] J. Xu *et al.*, "Forecast of dengue cases in 20 chinese cities based on the deep learning method," *International Journal of Environmental Research and Public Health*, vol. 17, no. 2, Jan. 2020, doi: 10.3390/ijerph17020453.

[43] A. Dairi, F. Harrou, A. Zeroual, M. M. Hittawe, and Y. Sun, "Comparative study of machine learning methods for COVID-19 transmission forecasting," *Journal of Biomedical Informatics*, vol. 118, 2021. doi: 10.1016/j.jbi.2021.103791.

[44] H. Saleh, E. Amer, T. Abuhmed, A. Ali, A. Al-Fuqaha, and S. El-Sappagh, "Computer aided progression detection model based on optimized deep LSTM ensemble model and the fusion of multivariate time series data," *Scientific Reports*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-42796-6.

[45] B. Yan, "An improved method for the fitting and prediction of the number of covid-19 confirmed cases based on LSTM," *Computers, Materials & Continua*, vol. 64, no. 3, pp. 1473–1490, 2020, doi: 10.32604/cmc.2020.011317.

[46] A. A. Ewees, M. A. A. Al-qaness, L. Abualigah, and M. A. Elaziz, "HBO-LSTM: Optimized long short term memory with heap-based optimizer for wind power forecasting," *Energy Conversion and Management*, vol. 268, Sep. 2022, doi: 10.1016/j.enconman.2022.116022.

[47] L. Zhou, C. Zhao, N. Liu, X. Yao, and Z. Cheng, "Improved LSTM-based deep learning model for COVID-19 prediction using optimized approach," *Engineering Applications of Artificial Intelligence*, vol. 122, Jun. 2023, doi: 10.1016/j.engappai.2023.106157.

[48] S. Hochreiter and J. U. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, doi: 10.1162/neco.1997.9.8.1735.

[49] S. Ghafouri-Fard, H. Mohammad-Rahimi, P. Motie, M. A. S. Minabi, M. Taheri, and S. Nateghinia, "Application of machine learning in the prediction of COVID-19 daily new cases: A scoping review," *Heliyon*, vol. 7, no. 10, 2021.

[50] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.

[51] M. A. Majeed, H. Z. M. Shafri, A. Wayayok, and Z. Zulkafli, "Prediction of dengue cases using the attention-based long short-term memory (LSTM) approach," *Geospatial Health*, vol. 18, no. 1, 2023, doi: 10.4081/gh.2023.1176.

BIOGRAPHIES OF AUTHORS

Eko Sediyono earned his professorship in 2008. He is a member of IEEE Computer Society #41605422. He completed his undergraduate studies in 1985 at the Bogor Institute of Agriculture with a major in Statistics. He received a master's degree in 1994 from the University of Indonesia and a doctorate in 2006 with a major in computer science. He is a lecturer and Vice-Rector of Research Innovation and Entrepreneurship at Satya Wacana Christian University, Indonesia. His research interests are data sciences, algorithms, and image processing. He can be contacted at email: eko@uksw.edu.

Sri Ngudi Wahyuni (Dr. Cand) is a senior lecturer at University of Amikom Yogyakarta Indonesia. She received her Dr. candidacy from the Satya Wacana Christian University of Indonesia in 2020, her bachelor's degree in engineering in 2005 from Ahmad Dahlan University, and her master's degree in Computer Science in 2014 from Indonesian Islamic University. She is currently working as a senior lecturer in the Informatics Management Study Program at the Faculty of Computer Science, University of Amikom Yogyakarta in Indonesia. Her research interest in data science, data mining, and computer science. She can be contacted at email: yuni@amikom.ac.id.

Irwan Sembiring completed his bachelor's degree in 2001 at UPN Veteran Yogyakarta, Indonesia. His master's degree was completed in 2004 from Gadjah Mada University Yogyakarta, Indonesia, and his doctorate was completed in 2016 at Gadjah Mada University Indonesia. His research interests are network security, information systems, and digital forensics. Now, he is a lecturer at the Faculty of Information Technology at Satya Wacana Christian University, Salatiga, Indonesia. He can be contacted at email: irwan@uksw.edu.