JOIN (Jurnal Online Informatika)
p-ISSN: 2528-1682, e-ISSN: 2527-9165
Volume 10 Number 1 | June 2025: 216-226
DOI: 10.15575/join.v10i1.1356

Simulation and Empirical Studies of Long Short-Term Memory Performance to Deal with Limited Data

Khusnia Nurul Khikmah[1], Kusman Sadik[2*], Khairil Anwar Notodiputro[3]
[1,2,3]Statistics Department, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, West Java, 16680 Indonesia

Article Info

Article history:

Received Aug 1, 2024
Revised Oct 28, 2024
Accepted Oct 28, 2024
Published June 05, 2025

Keywords:

Functional autoregressive
Long short-term memory
Multiple long short-term memory
Poverty
Simulation study

ABSTRACT

This research is proposed to determine the performance of time series machine learning in the presence of noise, where this approach is intended to forecast time series data. The approach method chosen is long short-term memory (LSTM), a development of recurrent neural network (RNN). Another problem is the availability of data, which is not limited to high-dimensional data but also limited data. Therefore, this study tests the performance of long short-term memory using simulated data, where the simulated data used in this study are data generated from the functional autoregressive (FAR) model and data generated from the functional autoregressive model of order 1 FAR(1) which is given additional noise. Simulation results show that the long short-term memory method in analyzing time series data in the presence of noise outperforms by 1-5% the method without noise and data with limited observations. The best performance of the method is determined by testing the analysis of variance against the mean absolute percentage error. In addition, the empirical data used in this study are the percentage of poverty, unemployment, and economic growth in Java. The method that has the best performance in analyzing each poverty data is used to forecast the data. The comparison result for the empirical data is that the M-LSTM method outperforms the LSTM in analyzing the poverty percentage data. The best method performance is determined based on the average value of the mean absolute percentage error of 1-10%.

Corresponding Author:

Kusman Sadik
Statistics Department, Faculty of Mathematics and Natural Sciences, IPB University
Meranti Street, Wing 22 Level 4. IPB Darmaga Campus, Bogor 16680 Indonesia.
Email: kusmans@apps.ipb.ac.id

1. INTRODUCTION

Statistical analysis research has developed rapidly and is used in various fields. This development encourages researchers to find the best method or approach for various problems. One of the results of this approach is used in forecasting. Forecasting is one method that can help make decisions based on past and present data [1]. This forecasting approach is developing rapidly, ranging from statistical methods such as autoregressive (AR) to the latest, namely deep learning. This development is motivated by the various types of data and data-related problems that exist in life. One problem in existing data requiring a renewed study of the approach method is time series data.

Time series data problems that depend on the data are on data with a limited amount and contain noise. The study of this problem in this research is proposed using a functional time series data approach using the functional time series method [2]. This approach is one of the relatively new statistical models with the influence of noise [3]. Functional time series is a statistical analysis method for time series data where the data variable underlying the analysis is a function. Time series data reconstructed into a function is done with the consideration that data that has become a smooth function allows the existence of time series data that is not stationary and is affected by the presence of noise in the model so that it is possible to be non-linear when generating data with the model. Based on these considerations, the use of models makes sense. The method in functional time series used in this study is functional autoregressive (FAR) [2],[3],[4].

A recent approach whose studies need to be compared clearly and precisely to forecast this functional data is time series machine learning [5]. Previous research by [6] mentioned that time machine learning has good forecasting capabilities. One of the development approaches in machine learning is artificial neural networks, and the development method is long short-term memory. In general, the long short-term memory method was chosen in this study with the consideration that long short-term memory can recognize data patterns very well and can forecast time series data well, besides previous research [7], [8], [9] concluded that the long short-term memory (LSTM) method has good accuracy and forecasting results.

Based on the described description, this research is intended to examine the advantages of forecasting results related to the long short-term memory method, which in previous studies was mentioned to be able to forecast well on high-dimensional data and data containing noise [10]. In addition, this research also provides a view of the actual problems and availability of data with high-dimensional data and limited data with a deep learning approach. Therefore, this research generates data with various scenarios of the amount of data and noise with the long short-term memory method. On the other hand, poverty is one of the problems related to time series data that we often encounter in the socio-population field in everyday life. There are several adverse effects of poverty on people's lives and a country's economic situation. Therefore, this research will also compare and contrast the LSTM and M-LSTM approaches and apply them to classified district and city-level poverty data in Java.

2. METHOD

2.1. Autoregressive (AR)

Autoregressive is one of the linear prediction modelling techniques in statistics [3]. The autoregressive model uses a function of previous values, a particular form of time series, as a prediction. This value is symbolized as p, which states the number of prior values used to predict the current value. Suppose that $Y_t it$ is a stationary time series. The autoregressive model with order p AR(p) is mathematically defined as [11]:

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \quad (1)$$

Where μ is an intercept constant, ϕ_{i} where $i = 1,2,3, \ldots, p$ are the model parameters to be estimated, Y_{t-i} where $i = 1,2,3, \ldots, p$ are the previous time series values, and e_t is the residual error where $e_t \sim WN(0, \sigma_e^2)$. The parameters of the autoregressive model ϕ_{i} can be estimated using several techniques such as Yule-Walker ([12], least square [13], [14], maximum likelihood [15]. In autoregressive methods, the model chosen to be used in the analysis is strongly influenced by the lag selection of the partial autocorrelation function (PACF) plot, where the best model will be chosen based on the most significant accuracy value of the tentative model. This accuracy can be calculated, for example, using Akaike's information criterion (AIC) [16], [17], [18], Bayesian information criterion (BIC) [19], [20].

2.2. Functional Data Analysis (FDA)

Functional data analysis is one of the statistical analysis methods where the data of the variable underlying the study is a function [21]. This method is also a relatively new one for analysis in statistics [3]. It has several advantages, including being able to analyse very heterogeneous data, being able to forecast data better than standard forecasting methods, being able to handle data with variables that have varying time relationships, having relatively low errors compared to traditional forecasting methods, and being able to capture data with complex patterns [22]. Functional data analysis is used to analyze data by assuming that each data is a single structured functional object, which can be used for time series data in the development of research. This data is a smooth curve that can be in the form of time or space. Suppose the data used is time-dependent $t_1, t_2, \dots, t_n$ defined as $Y_{i,t_1}, Y_{i,t_2}, \dots, Y_{i,t_n}$, then to reconstruct the data into a function, one of the methods that can be used is the basis function [2]. The basis function is a set of special functions ϕ_i from the functional space. Suppose ϕ_i where $i = 1,2,3, \dots, N$ is the basis function of the functional area, then the basis expansion $Y_t(k)$ is mathematically defined as follows:

$$Y_t(k) = \sum_{i=1}^{N} \alpha_i \phi_i(k) \tag{2}$$

With α_i where $i = 1,2,3, \dots, N$ are actual values of the coefficients [23]. Some of the bases that can be used are spline basis [24], Fourier basis [25], and wavelet basis [26]. Data reconstructed into this function can then be further analysed, one of which is by using functional autoregressive.

2.3. Functional Autoregressive (FAR)

Functional autoregressive is one of the methods in statistics where the resulting model [2]. In its development, this method is widely used to analyse time series problems in various fields, such as energy, economics, and climate. Suppose the function $Y_t(k)$ has a mean function $\mu(k)$ and covariance $C(x)$, namely:

$$C(x) = \int_0^1 E\big(Y_t(\pi), Y_t(k)\big), x(k)dk \tag{3}$$

By using Mercer's theorem that the condition that a function can be a kernel function must produce a kernel matrix that is positive-semi-definite [27], then:

$$C(x) = \sum_{j=1}^{\infty} \lambda_j \langle v_j, x \rangle v_j \,, j \in \mathbb{N} \tag{4}$$

Where λ_j is the eigenvalue in descending order and v_j is the normalized eigen function.

$$C(v_j) = \lambda_j v_j \tag{5}$$

And

$$\|v_j\| = 1 \tag{6}$$

The Karhonen-Loève theorem states that if $v_j(k)$ is an orthonormal basis of $L^2[0,1]$, then $\langle Y_t - \mu, v_j \rangle$ is the principal component functional of $Y_t(k)$ and the value of $Y_t(k)$ can be expressed as follows:

$$Y_t(k) = \mu(k) + \sum_{j=1}^{\infty} \langle Y_t - \mu, v_j \rangle v_j(k) \tag{7}$$

The estimator of the functional time series parameter $Y_t(k), t = 1,2,3, \dots, n$ for the mean $\mu(k)$ is:

$$\hat{\mu}(k) = \frac{1}{n} \sum_{t=1}^{n} Y_t(k) \tag{8}$$

and the covariance $C(x)$ is:

$$\hat{C}(x) = \frac{1}{n} \sum_{t=1}^{n} \langle Y_t(k) - \hat{\mu}(k), x \rangle \big(Y_t(k) - \hat{\mu}(k)\big) \tag{9}$$

where $x \in H$. Thus, the functional autoregressive model with order p (FAR(p)) is defined as follows:

$$Y_t(k) - \mu(k) = \sum_{i=1}^{p} \psi_i \left(Y_{t-i}(k) - \mu(k)\right) + e_t(k) \tag{10}$$

where ψ_i is a bounded linear operator where $\psi_i : H \to H$, Y_{t-i} is the i-th lag of the Y_t curve, and $e_t(k)$ is white noise with mean equal to 0 and $e_t(k) \in L_H^2$ [2].

2.4. Long Short-Term Memory (LSTM)

One of the statistical machine-learning forecasting techniques, long short-term memory (LSTM), is derived from recurrent neural networks (RNN) [5]. This method frequently studies time series data problems in various sectors, including health, economics, and climate, because it provides better forecasts than typical time series prediction methods. The system's long short-term memory (LSTM) features a multi-layered architectural design. The top layer is the input layer, followed by the hidden layer and the output layer. This long short-term memory (LSTM) has one memory cell in the buried layer, but several gates are included within one memory cell. Figure 1 depicts the long short-term memory's (LSTM) overall architecture.

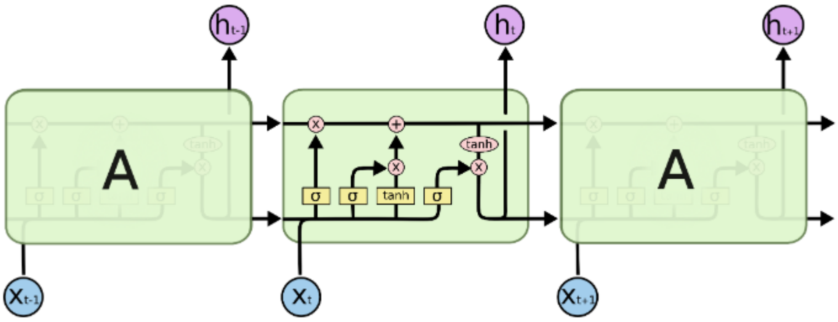


Figure 1. General architecture of long short-term memory (LSTM) based on [28].

The input gate is a gate that accepts both fresh inputs and prior outputs. The gate in this procedure returns a value of 0 or 1. Assume that the input value in_t is specified as follows:

$$in_t = \sigma(W_i Y_t + W_i h_{t-1} + b_i) \quad (11)$$

Suppose the memory cell's potential value, designated as $\bar{C_t}$, is as follows:

$$\bar{C_t} = \tanh(W_c Y_t + W_c h_{t-1} + b_c) \quad (12)$$

Where h_{t-1} is the state at time $t-1$, b is the bias, and W is the input gate weight. The forget gate utilizes an activation function, receiving input at time t and output at time $t-1$. Assume that the forget gate f_t is defined as follows:

$$f_t = \sigma\left(W_f Y_t + W_f h_{t-1} + b_f\right) \quad (13)$$

As a result, if the symbol C_t represents the updated memory cell's state, then:

$$C_t = in_t * \bar{C_t} + f_t * C_{t-1} \quad (14)$$

The gate that regulates how many states travel through this gate is the output gate. Let o_t represent the output gate's value, which is specified as follows:

$$out_t = \sigma(W_o Y_t + W_o h_{t-1} + V_o C_t + b_o) \quad (15)$$

If h_t is the cell's ultimate output value, then:

$$h_t = out_t * \tanh(C_t) \quad (16)$$

The output gate value is out_t, and the updated memory cell state is C_t [5], [6]. Additionally, the sigmoid and tanh activation functions are methodically defined as follows:

$$\sigma = \frac{1}{1 + e^{-x}} \quad (17)$$

$$\tanh(x) = 2\sigma(2x) - 1 \quad (18)$$

Where x is the input data, sigmoid activation function value, $tanh$ activation function value, and exponential value e are all present [29]. Each layer's output is produced in long short-term memory using the activation function itself.

2.5. Multiple-Long Short-Term Memory (M-LSTM)

Multiple long short-term memory (M-LSTM) is a statistical machine learning prediction approach that is an upgraded long short-term memory method that uses multivariate input. In general, the multiple long short-term memory approach has the same architecture as long short-term memory, with the obvious distinction being the variable input component [30], [31].

2.6. Analysis Procedure

1. The following analysis steps for simulation study with LSTM are broadly divided into three stages, namely:
   a. Generating data, first determining the number of observations and parameter values of simulation data to be generated according to the scenario, then generate data with FAR(1) with two predetermined scenarios (D_1: $FAR(1)$ with $e(k)\sim N(0.5, 1)$ and D_2: $FAR(1)$ from the generated data is taken 10% randomly and given noise $e(k)\sim N(0.5,\ 10)$, which is added to the initial data to form new data, namely $FAR^*(1)$). Third, divide the data into training data and testing data.
   b. Modelling and forecasting by using the LSTM to the generated training data, then doing forecasting and calculating accuracy
   c. Model performance evaluation, first repeating steps A and B 100 times, then evaluating LSTM based on the value of the calculated MAPE error measure. The last, perform hypothesis testing using ANOVA, where: H_0: $\mu_1 = \mu_2$ (the MAPE value of long short-term memory D_1 is equal to D_2) and H_1: $\mu_1 < \mu_2$ (the MAPE value of long short-term memory D_1 is smaller than D_2).

2. The empirical study data analysis procedure using LSTM and M-LSTM as follows:
   First, pre-process the data to tidy up the structure of the data used in the study, then explore the data to see the visual of the data pattern in general. The third is to conduct the stationary test, test the homogeneity of variance, see the seasonal effect of the data and divide the data into training data and test data. Analyse the data with LSTM and M-LSTM methods, doing hypothesis testing using ANOVA and lastly, perform forecasting and interpretation of analysis results.

3. RESULT AND DISCUSSION

Based on the research objectives, this section is divided into two studies: simulation and empirical. The simulation study is intended to determine the performance of the approach based on data generated with the specified scenario to examine the performance of LSTM in limited or no data and the influence of noise. In comparison, empirical studies are carried out with the aim of actual application of generalizations made in simulation studies where the empirical data used is data with a limited amount and contains noise.

3.1. Simulation Study

The simulation data for this study will be generated using the following scenarios. First, the data is generated from the functional autoregressive model with order 1 ($FAR(1)$) with $e(k) = N(0.5, 1)$. Second, the data is generated from the functional autoregressive model with order 1 ($FAR(1)$) from the generated data is taken 10% randomly and given noise $e(k)\sim N(0.5,\ 10)$, which is added to the initial data to form new data, namely $FAR^*(1)$. So, the performance of the long short-term memory method when applied to data generated with functional autoregressive model of order 1 ($FAR(1)$) with $e(k) = N(0.5, 1)$ or functional autoregressive model of order 1 ($FAR(1)$) with the generated data taken 10% randomly and given noise $e(k)\sim N(0.5,\ 10)$ which is added to the initial data to form new data, namely $FAR^*(1)$. The first stage is to generate data using a functional autoregressive model with order 1 ($FAR(1)$). functional autoregressive model with order 1 ($FAR(1)$):

$$Y_{t+1}(k) = \int_0^1 \psi(k,s)Y_t(s)ds + e_{t+1}(k)\,, t = 1,2,3, \dots, n \tag{19}$$

Then, it will be carried out in several study scenarios as shows at Table 1.

Table 1. $FAR(1)$ Simulation Scenarios $\psi = 0.5$

Scenario	Number of functional observations
1	30
2	350
3	400 (as per in [2])
4	450

The second stage, namely the generated data that has been following each generation scenario, will be repeated 100 times until 100 data with noise $e(k) \sim N(0.5, 1)$ and 100 data with 10% of the amount of data generated is taken randomly and given noise $e(k) \sim N(0.5, 10)$ which will then be used to compare the accuracy of the long short-term memory method on data containing noise and not having noise. In general, the data generated using the model in Table 1 with scenario 1 $FAR(1)$ with $e(k) \sim N(0.5, 1)$ has the visualization presented in Figure. 2. By repeating 100 times in each scenario, the data observations generated from the same model have different fluctuation ranges and are displayed with different colors. The results of the fluctuation range of data generated by model with 400 observations show a more extended fluctuation range than data with 30, 350, and 450 observations.

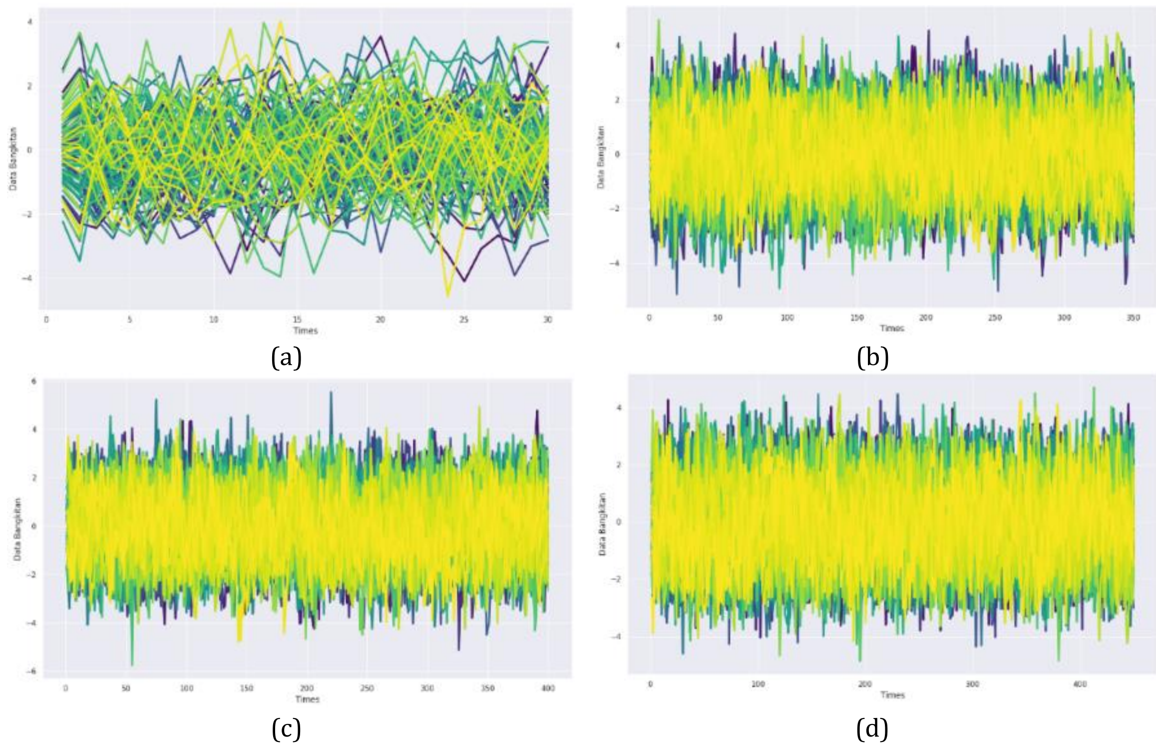
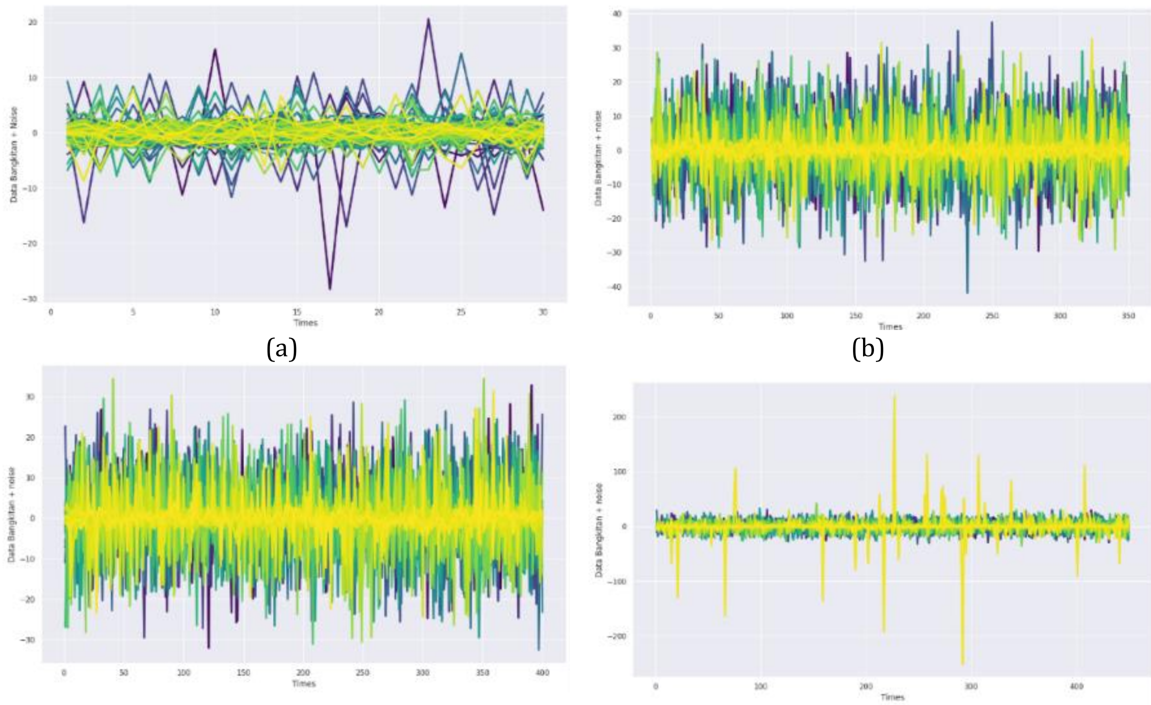


Figure 2. Scenario 1 generation data. (a) 30 observations. (b) 350 observations. (c) 400 observations. (d) 450 observations

The second scenario for generating data in this study, namely by using the $FAR(1)$ model of the generated data, is taken 10% randomly and given noise $e(k) \sim N(0.5, 10)$, which is added to the initial data to form new data, namely $FAR^*(1)$, has a visualization presented in Figure 3, by repeating 100 times in each observation scenario, the data generated from the same model has a different fluctuation range and is displayed in a different colour. The results of the fluctuation range of data generated by model with 450 observations show a more extended fluctuation range than data with 30, 350, and 400 observations.



(a) (b) (c) (d)

Figure 3. Scenario 2 generation data. (a) 30 observations. (b) 350 observations. (c) 400 observations. (d) 450 observations

The generation data in Figure 2 and Figure 3 are then subjected to modelling and forecasting. The modelling carried out is by using the long short-term memory method. The modelling results in each data scenario are obtained by comparing the modelling accuracy based on the RMSE value. The combination of parameters of the modelling results is carried out hypothesis testing, this test using factorial. This test is intended to detect the RMSE value of the level of each factor (number of observations, neurons in the hidden layer, and epoch) and the interaction between the elements.

Table 2. ANOVA p-value of modelling test results with LSTM

Source of variance	Scenario 1	Scenario 2
Observation	$< 2.00 \times 10^{-16}$	$< 2.00 \times 10^{-16}$
Number of neurons	$< 2.00 \times 10^{-16}$	$< 2.00 \times 10^{-16}$
Epoch	2.78×10^{-8}	0.879×10^{-1}
Interaction between observation and number of neurons	$< 2.00 \times 10^{-16}$	$< 2.00 \times 10^{-16}$
Interaction between observation and epoch	$< 2.00 \times 10^{-16}$	$< 2.00 \times 10^{-16}$
Interaction between number of neurons and epoch	$< 2.00 \times 10^{-16}$	$< 2.00 \times 10^{-16}$
Interaction between observation, number of neurons, and epoch	$< 2.00 \times 10^{-16}$	3.16×10^{-13}

The results of hypothesis testing show that all factors and their interactions have a real influence, which is established based on a significance level of 5%. It is known that the $p - value$ generated based on the analysis is < 0.05. Overall, the data model with a parameter combination of the number of neurons in the hidden layer of 10 and the number of epochs 50 has the smallest average value of RMSE distribution. The results of modelling carried out using two scenarios of data generation are then carried out forecasting. The accuracy of each design is selected based on the MAPE value, where the chosen results have been tested. The results of the hypothesis testing conducted show that all factors and their interactions have a real influence. Furthermore, this study tested the hypothesis on the data scenarios presented in Table 3.

Table 3. ANOVA p-value of forecasting test results with LSTM

Source of variance	Scenario 1	Scenario 2
Observation	$< 2.00 \times 10^{-16}$	$< 2.00 \times 10^{-16}$
Number of neurons	$< 2.00 \times 10^{-16}$	$< 2.00 \times 10^{-16}$
Epoch	2.76×10^{-7}	0.857×10^{-3}
Interaction between observation and number of neurons	$< 2.00 \times 10^{-16}$	$< 2.00 \times 10^{-16}$
Interaction between observation and epoch	$< 2.00 \times 10^{-16}$	$< 2.00 \times 10^{-16}$
Interaction between number of neurons and epoch	$< 2.00 \times 10^{-16}$	$< 2.00 \times 10^{-16}$
Interaction between observation, number of neurons, and epoch	$< 2.00 \times 10^{-16}$	$< 2.00 \times 10^{-16}$

Table 4 shows the RMSE and MAPE values with parameter combinations in each data scenario based on the hypothesis testing results.

Table 4. RMSE and MAPE value of forecasting test results with LSTM

Number of observations	Neurons in hidden layer	Epoch	Average RMSE		Average MAPE	
			Scenario 1	Scenario 2	Scenario 1	Scenario 2
30	5	50	0.3686	0.3470	51.21	48.33
		100	0.1573	0.1795	24.08	22.84
	10	**50**	**0.0956**	**0.0928**	**12.27**	**11.60**
		100	0.1005	0.0960	12.73	12.51
350	5	50	0.3735	0.4149	39.06	39.50
		100	0.3451	0.4059	39.59	38.19
	10	**50**	**0.3432**	**0.3274**	**33.30**	**31.52**
		100	0.3445	0.3639	37.18	34.33
400	5	50	0.4130	0.4245	42.21	41.12
		100	0.4017	0.4150	42.73	40.36
	10	**50**	**0.3373**	**0.3352**	**34.78**	**32.45**
		100	0.3737	0.3824	38.10	35.95
450	5	50	0.4433	0.4388	38.05	40.84
		100	0.4466	0.4371	39.69	41.79
	10	**50**	**0.3437**	**0.3366**	**35.39**	**30.45**
		100	0.3933	0.3937	36.83	40.69

Hypothesis testing is carried out on the data scenario. Hypothesis testing of significant differences in forecast accuracy between data generated without noise and noise (Table 5).

Table 5. Hypothesis testing result of significant differences in forecast accuracy

Source of variance	$p-value$	Source of variance	$p-value$
Scenario	2.72×10^{-4}	Interaction between observation and epoch	$< 2.00 \times 10^{-16}$
Observation	$< 2.00 \times 10^{-16}$	Interaction between number of neurons and epoch	$< 2.00 \times 10^{-16}$
Number of neurons	$< 2.00 \times 10^{-16}$	Interaction between scenario, observation, and number of neurons	9.64×10^{-3}
Epoch	1.53×10^{-9}	Interaction between scenario, observation, and epoch	6.87×10^{-4}
Interaction between scenario and observation	2.18×10^{-4}	Interaction between scenario, number of neurons, and epoch	9.27×10^{-3}
Interaction between scenario and number of neurons	6.63×10^{-2}	Interaction between observation, number of neurons, and epoch	$< 2.00 \times 10^{-16}$
Interaction between scenario and epoch	1.25×10^{-1}	Interaction between scenario, observation, number of neurons, and epoch	1.01×10^{-3}
Interaction between observation and number of neurons	$< 2.00 \times 10^{-16}$	Interaction between scenario, observation, and number of neurons	9.64×10^{-3}

The results of modelling and forecasting carried out using two scenarios of generated data show that with the best parameter combination scenario, both are shown in data with the second scenario or data developed with the $FAR(1)$ model with 10% of the amount of data generated randomly and given noise $e(k) \sim N(0.5, 10)$. Following previous research [32], the study found that the long short-term memory method is suitable and has good accuracy for data containing noise. In addition, random retrieval of data or sub-samples is based on previous research [33], mentioning that it can increase precision. Another study by [34] noted that applying sub-samples can reduce overfitting. The details of these differences are presented in Table 6.

Table 6. Comparison results of forecasting with the best parameter combination of LSTM

Accuracy	Scenario 1	Scenario 2	$p - value$
Average MAPE	12.27	11.60	6.22×10^{-3}
	33.30	31.52	5.04×10^{-4}
	34.78	32.45	3.78×10^{-4}
	35.39	30.45	1.93×10^{-4}

Furthermore, this study conducted hypothesis testing on the significant difference in forecast accuracy between data generated without noise and with noise. Further testing is done using the Z test. The Z test results show a substantial difference in the average MAPE of the two data with the best parameter combination. This is known based on the test results showing that all $p - values$ are less than 0.05. Based on hypothesis testing, it is proven that there is a significant difference in the average MAPE of the forecasting results so that it can be obtained that the LSTM method has better performance in scenario 2 based on Table 6 or on data with additional noise.

This study took one data in the best scenario for visualization, namely the data with the best forecasting accuracy on data generated with the $FAR(1)$ model with 10% of the total data generated randomly taken and given noise $e(k) \sim N(0.5, 10)$ with a parameter combination of the number of neurons in the hidden layer of 10 and epoch 50. The results of this visualization are presented in Figure 4.

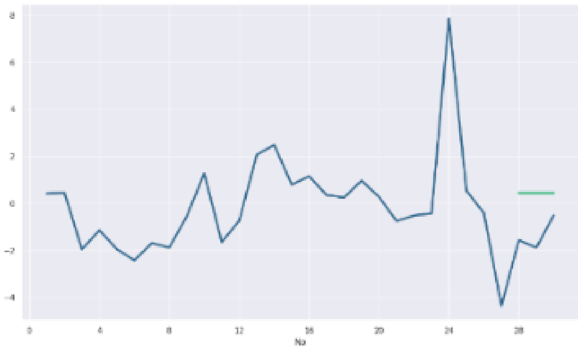


Figure 4. Forecasting simulated data of 30 observations of the best parameter combination

3.2. Empirical Study

The empirical data used in this study is secondary data taken from the official website of the Central Bureau of Statistics of six provinces in Java or 119 regencies/cities where the data used is taken from 1980 to 2021. The data on the percentage of poverty at the district/city level is then divided into three groups according to the poverty percentage variable, namely $\leq q_1$; (q_1, q_3); $\geq q_3$, where $q_1 = 8.085$ and $q_3 = 17.290$. Furthermore, the categorized data is divided into 90 percent as training data and 10 percent as testing data, where the data used as input is Y_{t-1}, the percentage of poverty at time $t-1$. $X_{1_{t-1}}$ unemployment at time $t-1$. $X_{2_{t-1}}$ economic growth at the district/city level at time $t-1$. Meanwhile, the desired output is Y_t the percentage of poverty at the regencies/cities level at time t. This research begins with data collection, and then pre-processing of this data begins with checking for missing data, which is handled using the mean imputation method. Next, the parameter initialization stage is carried out to model the training data (pre-processed data). This research uses one long short-term memory layer with 16 and 32 neurons and one dense layer. The optimizer used is adaptive moment estimation (Adam) with an epoch of 40 and a number of batch sizes 64. [35] state that the forecasting results generated from a long short-term memory model are influenced by the number of neuron units that build the model. The neuron units used in model formation are selected based on the value of multiples of 16. This study uses 16 and 32. The analysis results show that forecasting with data division, in general, has outstanding accuracy on data with the $\leq q_1$ category, where the best method is M-LSTM with a neuron count of 16. The best accuracy results are the same as the analysis carried out in each district and city.

Table 7. MAPE of LSTM and M-LSTM models on data divided into 3 groups

Category	LSTM Neuron 16	M-LSTM Neuron 16	LSTM Neuron 32	M-LSTM Neuron 32
$\leq q_1$	4.2827	4.1793	4.4156	7.4180
(q_1, q_3)	12.2659	13.9213	40.3223	31.7204
$\geq q_3$	25.8223	28.8959	35.6499	39.3727

Simulation and empirical studies carried out in this study show the same conclusion that the LSTM method has good performance on data with a limited amount and has noise based on the MAPE value of forecasting results and hypothesis testing conducted on testing data. The simulation studies' results by adding noise to the generated data also show that the LSTM method has a good forecasting MAPE value (Table 7). Following the research objectives the LSTM method is well used as a time series data approach with a limited amount and data containing noise.

4. CONCLUSION

The results of research conducted on simulation studies, namely data generated with the $FAR(1)$ model, show that in determining the best parameters for generation data, both influenced by noise and not, it is obtained that the LSTM method parameter with the number of neurons in the hidden layer of 10 and the number of epochs of 50 is a parameter with better performance compared to other parameter combinations that have been initiated at the beginning, where the MAPE of the forecast that has been tested by the analysis of variance of the generation data given noise is superior to the generation data that is not given noise, which is 1-5%. While in the empirical study, namely the use of LSTM and M-LSTM, shows that the M-LSTM forecasting results are better than LSTM. This is indicated by the difference in the average value of MAPE from the two models with the MAPE forecast that the analysis of variance from M-LSTM has tested smaller by 1-10%. Empirically, the M-LSTM method is well used to forecast the percentage of poverty in districts and cities in Java.

REFERENCES

[1] F. Petropoulos *et al.*, "Forecasting: theory and practice," *Int J Forecast*, 2022.

[2] S. Jiao, A. Aue, and H. Ombao, "Functional Time Series Prediction Under Partial Observation of the Future Curve," *J Am Stat Assoc*, pp. 1–29, May 2021, doi: 10.1080/01621459.2021.1929248.

[3] I. Shah, H. Bibi, S. Ali, L. Wang, and Z. Yue, "Forecasting one-day-ahead electricity prices for italian electricity market using parametric and nonparametric approaches," *IEEE Access*, vol. 8, pp. 123104–123113, 2020.

[4] Y. Chen, T. Koch, X. Xu, K. Lim, and N. Zakiyeva, "A review study of functional autoregressive models with application to energy forecasting," *WIREs Computational Statistics*, vol. 13, Jul. 2020, doi: 10.1002/wics.1525.

[5] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos Solitons Fractals*, vol. 135, p. 109864, 2020.

[6] H. Chung and K. Shin, "Genetic algorithm-optimized long short-term memory network for stock market prediction," *Sustainability*, vol. 10, no. 10, p. 3765, 2018.

[7] S. Bouktif, A. Fiaz, A. Ouni, and M. A. Serhani, "Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches," *Energies (Basel)*, vol. 11, no. 7, p. 1636, 2018.

[8] A. Sagheer and M. Kotb, "Time series forecasting of petroleum production using deep LSTM recurrent networks," *Neurocomputing*, vol. 323, pp. 203–213, 2019.

[9] M. Elsaraiti and A. Merabet, "A comparative analysis of the arima and lstm predictive models and their effectiveness for predicting wind speed," *Energies (Basel)*, vol. 14, no. 20, p. 6782, 2021.

[10] Z. Liu, X. Hu, L. Xu, W. Wang, and F. M. Ghannouchi, "Low computational complexity digital predistortion based on convolutional neural network for wideband power amplifiers," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 3, pp. 1702–1706, 2021.

[11] S. I. Alzahrani, I. A. Aljamaan, and E. A. Al-Fakih, "Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions," *J Infect Public Health*, vol. 13, no. 7, pp. 914–919, 2020, doi: https://doi.org/10.1016/j.jiph.2020.06.001.

[12] W. Żuławiński and A. Wyłomańska, "New estimation method for periodic autoregressive time series of order 1 with additive noise," *Int J Adv Eng Sci Appl Math*, vol. 13, no. 2, pp. 163–176, 2021, doi: 10.1007/s12572-021-00302-z.

[13] F. Ding, Y. Wang, J. Dai, Q. Li, and Q. Chen, "A recursive least squares parameter estimation algorithm for output nonlinear autoregressive systems using the input–output data filtering," *J Franklin Inst*, vol. 354, no. 15, pp. 6938–6955, 2017.

[14] F. Ding, D. Meng, J. Dai, Q. Li, A. Alsaedi, and T. Hayat, "Least Squares based Iterative Parameter Estimation Algorithm for Stochastic Dynamical Systems with ARMA Noise Using the Model Equivalence," *Int J Control Autom Syst*, vol. 16, no. 2, pp. 630–639, 2018, doi: 10.1007/s12555-017-0001-x.

p-ISSN: 2528-1682
e-ISSN: 2527-9165

[15] J. Li and J. Zhang, "Maximum likelihood identification of dual-rate Hammerstein output-error moving average system," *IET Control Theory & Applications*, vol. 14, no. 8, pp. 1089–1101, 2020.

[16] J. E. Cavanaugh and A. A. Neath, "The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements," *Wiley Interdiscip Rev Comput Stat*, vol. 11, no. 3, p. e1460, 2019.

[17] H. de-G. Acquah, "Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship," *J Dev Agric Econ*, vol. 2, no. 1, pp. 1–6, 2010.

[18] D. R. Anderson, K. P. Burnham, and G. C. White, "Comparison of Akaike information criterion and consistent Akaike information criterion for model selection and statistical inference from capture-recapture studies," *J Appl Stat*, vol. 25, no. 2, pp. 263–282, 1998.

[19] A. G. Salman and B. Kanigoro, "Visibility forecasting using autoregressive integrated moving average (ARIMA) models," *Procedia Comput Sci*, vol. 179, pp. 252–259, 2021.

[20] Z. Zhao, Y. Liu, and C. Peng, "Variable selection in generalized random coefficient autoregressive models," *J Inequal Appl*, vol. 2018, no. 1, pp. 1–14, 2018.

[21] A. R. Ghumman, A.-R. Ateeq-ur-Rauf, H. Haider, and M. Shafiquzamman, "Functional data analysis of models for predicting temperature and precipitation under climate change scenarios," *Journal of Water and Climate Change*, vol. 11, no. 4, pp. 1748–1765, 2020.

[22] M. Dass and C. Shropshire, "Introducing Functional Data Analysis to Managerial Science," *Organ Res Methods*, vol. 15, pp. 693–721, Oct. 2012, doi: 10.1177/1094428112457830.

[23] A. Leroy, A. Marc, O. Dupas, J. L. Rey, and S. Gey, "Functional data analysis in sport science: Example of swimmers' progression curves clustering," *Applied Sciences*, vol. 8, no. 10, p. 1766, 2018.

[24] J.-L. Wang, J.-M. Chiou, and H.-G. Müller, "Functional data analysis," *Annu Rev Stat Appl*, vol. 3, pp. 257–295, 2016.

[25] P. Kokoszka and M. Reimherr, *Introduction to functional data analysis*. Chapman and Hall/CRC, 2017.

[26] M. Giacofci, S. Lambert-Lacroix, G. Marot, and F. Picard, "Wavelet-based clustering for mixed-effects functional models in high dimension," *Biometrics*, vol. 69, no. 1, pp. 31–40, 2013.

[27] H. Sun, "Mercer theorem for RKHS on noncompact sets," *J Complex*, vol. 21, no. 3, pp. 337–349, 2005, doi: https://doi.org/10.1016/j.jco.2004.09.002.

[28] C. Olah, "Understanding LSTM Networks," https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[29] B. Karlik and A. V. Olgac, "Performance analysis of various activation functions in generalized MLP architectures of neural networks," *International Journal of Artificial Intelligence and Expert Systems*, vol. 1, no. 4, pp. 111–122, 2011.

[30] S. M. Miran, S. J. Nelson, D. Redd, and Q. Zeng-Treitler, "Using multivariate long short-term memory neural network to detect aberrant signals in health data for quality assurance," *Int J Med Inform*, vol. 147, p. 104368, 2021.

[31] X. Xu, X. Rui, Y. Fan, T. Yu, and Y. Ju, "A multivariate long short-term memory neural network for coalbed methane production forecasting," *Symmetry (Basel)*, vol. 12, no. 12, p. 2045, 2020.

[32] S. Liu, X. Liu, Q. Lyu, and F. Li, "Comprehensive system based on a DNN and LSTM for predicting sinter composition," *Appl Soft Comput*, vol. 95, p. 106574, 2020.

[33] A. S. Acharya, A. Prakash, P. Saxena, and A. Nigam, "Sampling: Why and how of it," *Indian Journal of Medical Specialties*, vol. 4, no. 2, pp. 330–333, 2013.

[34] J. Ranstam and J. A. Cook, "LASSO regression," *British Journal of Surgery*, vol. 105, no. 10, p. 1348, Sep. 2018, doi: 10.1002/bjs.10895.

[35] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, "LSTM network: a deep learning approach for short-term traffic forecast," *IET Intelligent Transport Systems*, vol. 11, no. 2, pp. 68–75, 2017.