Riwayat: Educational Journal of History and Humanities, 6 (3), 2023, pp. 2210-2218
P-ISSN: 2614-3917, E-ISSN: 2775-5037
DOI: https://doi.org/10.24815/jr.v6i3.3445

Ozone Gases Value Forecasting Using Encoder-Decoder LSTM Model

Ni Ketut Intan Rahayu[1]✉, Devi Fitrianah[2], Elvin[3], Tannuru Marthamurtadha[4]
[1,2,3,4]Bina Nusantara University, Jakarta, Indonesia
Correspondence Author: ni.rahayu002@binus.ac.id✉

Article history
Received : 2023-04-23
Accepted : 2023-07-25
Published : 2023-08-31

Keywords:
Encoder-Decoder, Environmental Monitoring, Ozone, Long-Short-Term memory (LSTM)

Abstract: Climate change, as one of the impacts of global warming, has several consequences for the sustainability of living beings on Earth. It is necessary to monitor the trend of climate change. One way to monitor seasonal patterns of change is by analyzing the ozone content in the air. In addition to being an indicator of climate change, predicting the ozone gas content in the air is important because ozone gas has a direct impact on living organisms. By predicting the ozone gas content, it is hoped that preventive measures can be taken to prevent the adverse effects of ozone gas in the air. In the case of predicting ozone gases, there may be certain patterns that only become apparent over time, such as seasonal variations or long-term trends. A model that can capture these long-term dependencies will be better equipped to accurately predict ozone gas levels in the future. In this experiment, we proposed the use of Encoder-Decoder LSTM to predict ozone gas values.

Abstrak: Perubahan iklim, sebagai salah satu dampak dari pemanasan global, memiliki beberapa konsekuensi bagi keberlanjutan makhluk hidup di Bumi. Penting untuk memantau tren perubahan iklim. Salah satu cara untuk memantau pola perubahan musiman adalah dengan menganalisis kandungan ozon di udara. Selain menjadi indikator perubahan iklim, memprediksi kandungan gas ozon di udara penting karena gas ozon memiliki dampak langsung pada organisme hidup. Dengan memprediksi kandungan gas ozon, diharapkan dapat diambil tindakan preventif untuk mencegah efek buruk gas ozon di udara. Dalam hal memprediksi gas ozon, mungkin ada pola tertentu yang hanya menjadi jelas seiring berjalannya waktu, seperti variasi musiman atau tren jangka panjang. Model yang dapat menangkap ketergantungan jangka panjang ini akan lebih baik dalam memprediksi tingkat gas ozon di masa depan. Dalam eksperimen ini, kami mengusulkan penggunaan Encoder-Decoder LSTM untuk memprediksi nilai gas ozon.

Available online at
http://jurnal.usk.ac.id/riwayat/

INTRODUCTION

Ozone is a gas that plays a crucial role in both the Earth's atmosphere and at ground level (Zoran et al., 2020). While ozone in the stratosphere provides protection by absorbing harmful ultraviolet (UV) radiation from the sun, at ground level, it can be a pollutant and a major component of smog. Because ozone gas plays a crucial role in the lives of living organisms on Earth, designing a model to predict ozone gas concentrations is also important to do. Ozone forecasting systems are crucial for protecting human health, managing air quality, preserving the environment, facilitating urban planning, and guiding policy decisions. By providing timely and accurate information about ozone concentrations, these systems assist in minimizing the adverse effects of ozone pollution and promoting sustainable development.

In terms of air quality forecasting, its models are divided into deterministic, statistical, and hybrid approaches (Kök et al., 2017). In order to predict an atmospheric

pollutant concentration, the deterministic approach uses physical and chemical mechanisms, while the statistical approach uses historical data as inputs to learn the pattern of changes in ozone concentration by relying on changes in predictor variables over time. Due to the linearity of the early statistical models (such as linear regression-based models), improved machine learning and deep learning models have been utilized to tackle the non-linearity issue of such data. Deep learning has gained significant attention in recent years, primarily due to its ability to automatically extract meaningful features from raw data. Unlike traditional machine learning algorithms, deep learning algorithms utilize multi-layer networks to automatically learn and identify relevant features, eliminating the need for manual feature engineering. This feature extraction process significantly reduces the manual effort required in preparing data for analysis and enhances the efficiency and effectiveness of the learning process (Ettouney et al., 2009). The multi-layer architecture in deep learning algorithms provides the possibility of extracting air quality features with complex characteristics. In addition, sequential deep models, such as RNN and LSTM, are capable of learning temporal dynamic patterns from a sequence of input parameters.

The proposed forecasting model uses encoder-decoder LSTM as the model. The target of the proposed model is to capture temporal dependencies and handle sequence-to-sequence prediction tasks and encoder-decoder LSTM architecture is well-suited for that. An encoder learns a vector representation of the input time-series and the decoder uses this representation to reconstruct the time-series (Jacoby et al., 2021). In this context, the encoder LSTM takes historical ozone concentration data as input and encodes it into a fixed-length representation, capturing the underlying patterns and trends in the data. The decoder LSTM then takes this encoded representation and generates future ozone concentration predictions. The LSTM's recurrent nature allows it to capture long-term dependencies and patterns in the ozone data, which is crucial for accurate forecasting. The encoder-decoder structure enables the model to learn the complex relationships between past ozone concentrations and future predictions, making it a suitable choice for ozone forecasting applications.

1. Related Work

In the last 10 years, researchers have developed several models to improve the accuracy of Ozone predictions (Coman et al., 2008). This section focuses on explaining previous studies related to the development of forecast models for Ozone gas concentration. Coman et al. predicted the concentration of ozone gas every hour using a statistical method called a recursive structure involving a cascade of 24 multilayer perceptrons (MLP) arranged so that each MLP feeds the next one. In the designed MLP model, two types of MLP modes, namely 'static' and 'dynamic,' were compared. The performance of both architectures slightly increases when using meteorological predictors and decreases gradually with the prediction lag (Abdul-Wahab & Al-Alawi, 2002). This research concludes that the ‘static’ model is slightly more effective than the ‘dynamic’ one for the urban site, at least for the first 8 h of the prediction horizon.

(Lyu et al., 2020) proposed a multi-step prediction model for gas concentration time series based on the ARMA model, the CHAOS model and the Encoder-Decoder model (single-sensor and multi-sensor). The Encoder-Decoder model provides high robustness in a multi-step prediction and can predict the gas concentration for five different time steps. Its prediction error is significantly lower than those of the ARIMA and the CHAOS models (Lyu et al., 2020). Ettouney et al. develop and validate a neural-based modeling methodology applicable to site-specific short- and medium-term ozone concentration forecasting. The model utilizes two feed forward artificial neural networks (FFNN). The newly developed model improves the prediction accuracy over the conventional method (Feed Forward Neural Network) (Ettouney et al., 2009).

Abdul-Wahab and Al-Alawi designed a model to predict tropospheric ozone

concentration using an artificial neural network (ANN). This study concludes that the neural network can be used in modeling and predicting the ground level concentrations of ozone. This study has indicated the potential of the neural network approach for capturing the non-linear interactions between ozone and other factors and for the identification of the relative importance of these factors (Abdul-Wahab & Al-Alawi, 2002). Pires et al. proposed the optimization of artificial neural network models through genetic algorithms (GAs) for surface ozone concentration forecasting. GAs were applied to define the activation function in the hidden layer and the number of hidden neurons (Pires et al., 2012). In threshold models, the variables selected by GAs to define the O3 regimes were temperature, CO and NO2 concentrations (Pires et al., 2012).

Capturing long-term dependencies in time series data in this experiment is Ozone gases data is important because it allows the model to understand the relationships and patterns between data points that occur over an extended period of time (Sadhukhan & Yadav, 2023). In many cases, these relationships and patterns may not be apparent in short-term dependencies or individual data points. By capturing these long-term dependencies, the model can make more accurate predictions and improve its overall performance in time series forecasting tasks. The identified research gap that will be discussed in this study is the need for a model that can capture long-term dependencies in time series data. which is essential for accurately predicting Ozone gas concentrations.

2. Long Short-Term Memory

Long Short-Term Memory Neural Network (LSTM) is a type of Recurrent Neural Network (RNN). LSTM has the ability to learn which data should be used or ignored. LSTM is widely used for processing text, videos, and time-series data. Since more previous information can affect the accuracy of the model, LSTM is a reasonable choice for usage. The LSTM module, called the repeating module, has four interacting neural network layers, as shown in Figure 2 of the LSTM Module Repetition.

The symbols π and Σ represent the elements of wise multiplication and addition, respectively. The merging operation is represented by the dot (•) symbol. The basic component of LSTM is the cell state, a line that runs from the memory of the previous block (S t-1) to the memory of the current block (St). This allows information to flow straight down. The network can determine how much previous information flows. It is controlled through the first layer (σ1). The operation performed by this layer is given.

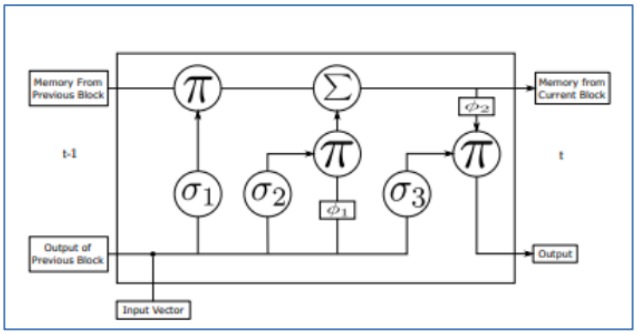
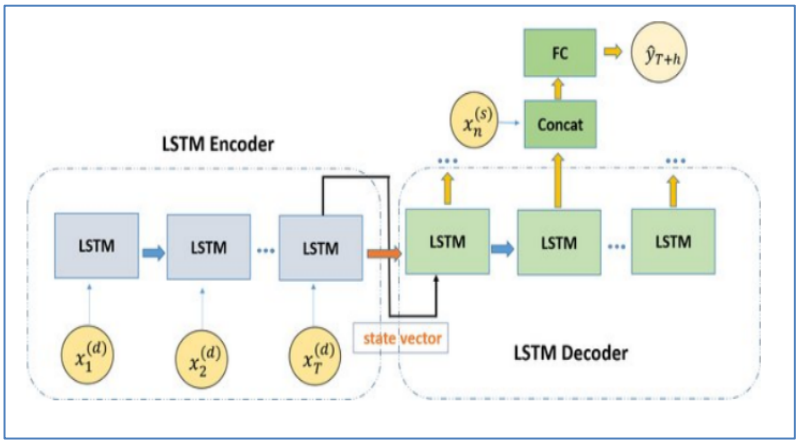


Figure 1. LSTM Architecture

3. Proposed Work

We propose an encoder-decoder LSTM model as a model for predicting the ozone gas content in numerical time units for the next one to three days. The use of the encoder-decoder architecture is because input sequence and the output sequence have different lengths and the entire input sequence needs to be summarized into a fixed-length context vector before generating the output sequence. The use of LSTM models in this study is because LSTM (Long Short-Term Memory) neural networks are commonly used in time series forecasting due to their ability to capture and learn complex patterns in sequential data. Traditional statistical time series forecasting methods often assume that the time series is linear and stationary and can therefore struggle to capture non-linear and non-stationary relationships in the data. LSTM networks, on the other hand, are capable of learning these more complex relationships by maintaining a memory of past inputs and outputs through a chain of memory cells. LSTM networks can also handle missing values and noisy data well by interpolating missing values and smoothing the time series data. In addition, LSTM networks can learn to model long-term dependencies in time series data by selectively forgetting or retaining information in their memory cells. This makes them especially useful for forecasting problems where there are time lags between inputs and outputs.

Figure 2. Encoder-Decoder LSTM Architecture



The proposed model:

a. Variable-Length Input and Output Sequences: Unlike traditional time series forecasting models, an Encoder-Decoder LSTM can handle variable-length input and output sequences
b. Contextual Information: The Encoder-Decoder architecture allows the model to maintain a summary or context vector of the input sequence. This context vector can then be used to generate the output sequence. By maintaining a summary of the input sequence, the model can capture the long-term dependencies in the time series data
c. Handling Non-Linear Relationships: An Encoder-Decoder LSTM can capture non-linear relationships between the input and output sequences. This is particularly useful for forecasting problems where the relationship between the input and output sequences is not linear, such as predicting energy demand
d. Multi-Step Forecasting: An Encoder-Decoder LSTM can be used to perform multi-step forecasting, where the model predicts multiple future time steps at once. In this experiment we will predict ozone gases in the next 24, 48 and 72 hours ahead.

METHODS

Experiments

Fig. 3 depicts the detailed workflow of this experiment. The experimental process begins with data collection and pre-processing phases, which will be described

in more detail in the next section. The next section will also explain the mechanism of data splitting and the model performance evaluation process. The model evaluation process also includes a performance comparison between the Encoder-Decoder LSTM model and one machine learning algorithm, Linear Regression.

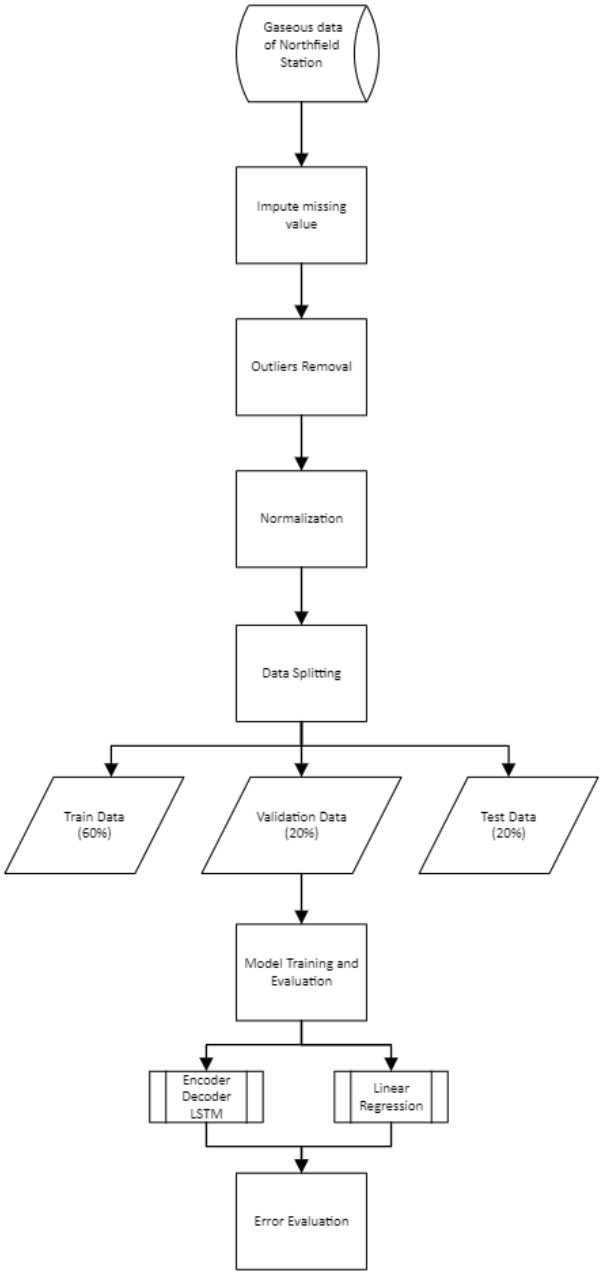


Figure 3 Detailed Workflow

1. Dataset Description And Data Preprocessing

In this experiment, we used hourly data on the ozone gas content from the Northfield Station with a data timeframe from January 2015 to December 2019. This data set contains validated gaseous pollution data for the for the Northeastern Adelaide region in monthly files. Information provided is Ozone values expressed in parts per million (ppm), Nitrogen oxides values expressed in parts per million (ppm), Sulfur dioxide values expressed in parts per million (ppm). There are 6 types of ozone gas content parameters used as input data parameters, namely: O3 UVA ppm, O3 4hr UVA ppm, NO Chemiluminescence ppm, NO2 calc Chemiluminescence ppm, NOx Chemiluminescence ppm, and SO2 UVF ppm. The encoder-decoder LSTM model will predict the values of the O3 UVA ppm and NO2 calc Chemiluminescence ppm parameters.

This experiment involved real data that required several data-preprocessing processes to improve the performance of the prediction model. The data pre-processing processes carried out in this study were missing value handling, normalization, and outlier removal. Missing values in the dataset were replaced with a value of 0. The normalization technique used in this experiment is MinMax Scaling Technique. Mathematical equation of MinMax normalization technique show as follow:

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Meanwhile, the outlier removal process used the capping technique, with a lower limit consisting of the first quartile value minus 1.5 times the interquartile value and an upper limit consisting of the third quartile value plus 1.5 times the interquartile value. The mathematical formula of the Capping technique is described as follows:

$$\boldsymbol{Lower\ Bound} = \boldsymbol{Q1} - (\mathbf{1.5} * \boldsymbol{IQR}) \quad (2)$$
$$\boldsymbol{Upper\ Bound} = \boldsymbol{Q3} + (\mathbf{1.5} * \boldsymbol{IQR}) \quad (3)$$

Where:

$$IQR = Q3 - Q1 \quad (4)$$

2. Model Training

After the data pre-processing process is applied to all input data, the next step is to divide the input sequences using the sliding window method. The size of the sliding window used in this study is 24. The model training process also includes searching for the optimal LSTM hyperparameter values. The optimal hyperparameter values were obtained using the Bayesian optimizer method. The number of optimum epochs is decided by using the early stopping method a maximum epoch limit of 100 and a patience value of 5. The maximum epoch value for predicting the O3 UVA ppm parameter is 15. The details number of hyperparameter of each model will be explained in the model comparison section. The results of the model training are illustrated in Fig. 4.

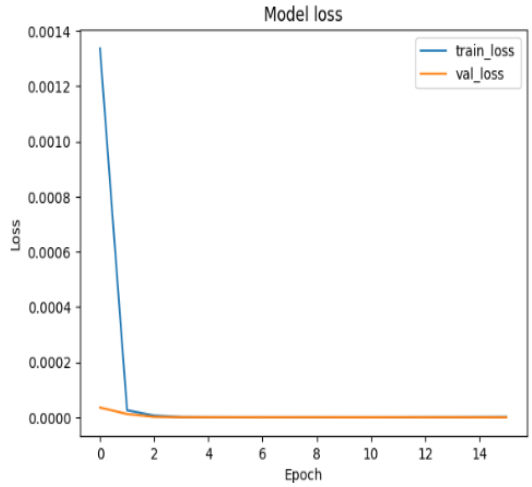


Figure 4 Training Curve of Encoder-Decoder LSTM Model

3. Model Evaluation

There are two different performance metrics that are used in this experiment: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). MAE measures the average absolute difference between the predicted and actual values, while RMSE measures the square root of the average squared difference between the predicted and actual values.

These metrics are useful because they provide an easy-to-interpret measure of the accuracy of the forecasting model. Additionally, both MAE and RMSE are sensitive to outliers in the data, which is important in many time series applications where outliers can have a significant impact on the performance of the model. The performance metrics are calculated as follows:

$$MAE = \frac{\sum_1^N |y_i - \hat{y}_i|}{N} \quad (3)$$

$$RMSE = \sqrt{\sum_1^N \frac{(y_i - \hat{y}_i)^2}{N}} \quad (4)$$

Where y_i is the actual value of the Ozone gas at time step i and y`_i is the predicted value of the Ozone gas at i^th time step, N is the sample size.

4. Model Comparison

The performance of the Encoder-Decoder LSTM model is compared with a model that uses one of the machine learning algorithms, Linear Regression. This comparison is made to evaluate whether the performance of the deep learning model using the Encoder-Decoder LSTM model is more effective or not compared to the Linear Regression machine learning model. Both types of models are used to predict the value of O3 UVA ppm and NO2 Chemiluminescence ppm. The models predict the values of both parameters for a period of 24, 48, and 72 hours in advance.

The hyperparameter values of the Encoder-Decoder LSTM model are presented in Table I.

Table 1 Hyperparameter of Encoder-Decoder LSTM Model

Param	Epoch	Batch Size	Hidden Dim	Learning rate	Dropout Rate	Optimizer
O3 UVA	15	96	64	0.01	0.0	Adam
NO2	16	112	160	0.001	0.0	Adam

RESULTS AND DISCUSSION

In this experiment, an encoder-decoder LSTM model is proposed to predict the values of O3 UVA and NO2 Chemiluminescence gases. The model is used to predict the values of both Ozone gas composition parameters for the next 24, 48, and 72 hours. The performance of the model is evaluated using the performance metrics of MAE and RMSE.

O3 Uva Prediction

Table II shows the RMSE and MAE values for each model to predict the O3 UVA gas concentration. From the table, it can be seen that the encoder-decoder LSTM model performs better than the Linear Regression model. The table also shows that there is no significant change in the RMSE and MAE values for predictions at different timesteps.

Table 2 Performance Evaluation Result Of O3 Uva Prediction

Model	Timesteps ahead	RMSE	MAE
Linear Regression	24 hours	0.00798	0.00615
	48 hours	0.00843	0.00652
	72 hours	0.00895	0.00692
Enc-Dec LSTM	24 hours	0.00166	0.00165
	48 hours	0.00166	0.00165
	72 hours	0.00166	0.00165

Predicted vs actual value of O3 UVA in 100 timesteps ahead by encoder-decoder LSTM depicted in fig.5 and the predicted vs actual value of O3 UVA in 100 timesteps ahead by Linear Regression depicted in figure 6.

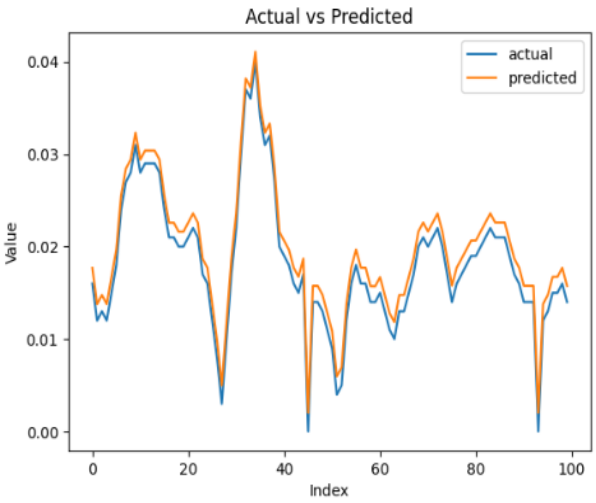


Figure 5. Actual vs Predicted Value of O3 UVA by Encoder-Decoder LSTM Model

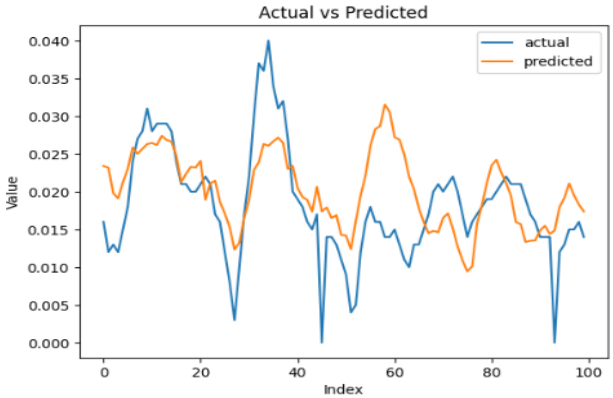


Figure 6. Actual vs Predicted Value of O3 UVA by Linear Regression Model

Nox Chemiluminescence Prediction

Table III shows the RMSE and MAE values of each model for predicting the NO2 Chemiluminescence gas content. From the table, it can be seen that the encoder-decoder LSTM model performs better than the Linear Regression model. The table also shows that there is no significant change in the RMSE and MAE values for predictions at different timesteps.

Table 3 Performance Evaluation Result Of No2 Chemiluminescence Prediction

Model	Timesteps ahead	RMSE	MAE
Linear Regression	24 hours	0.00433	0.00329
	48 hours	0.00449	0.00340
	72 hours	0.00454	0.00346
Enc-Dec LSTM	24 hours	0.00024	0.00013
	48 hours	0.00024	0.00013
	72 hours	0.00024	0.00013

Predicted vs actual value of NO2 calc Chemiluminescence in 100 timesteps ahead by encoder-decoder LSTM depicted in fig.8 and the predicted vs actual value of NO2 calc Chemiluminescence in 100 timesteps ahead by Linear Regression depicted in fig.7.

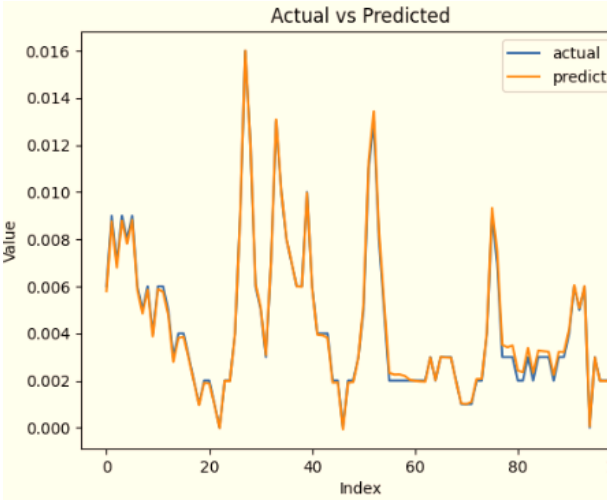


Figure 7. Actual vs Predicted Value of NO2 calc Chemiluminescence by Linear Regression Model

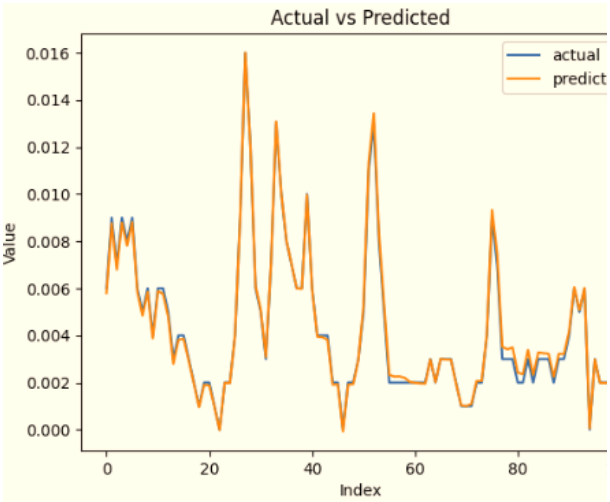


Figure 8. Actual vs Predicted Value of NO2 calc Chemiluminescence by Encoder-Decoder LSTM Model

CONCLUSION

This research discusses the appropriate model that can be used to predict the value of Ozone gas content in the North Eastern Adelaide region. To obtain contextual information from input data, the use of the encoder-decoder architecture with the LSTM algorithm is proposed. By obtaining this contextual information, the model can capture long-term dependencies in time series data.

From the results of this study, it can be concluded that the Encoder-Decoder LSTM model is effective as a predictor of Ozone gas content. This study also proves that the encoder-decoder LSTM model has better model performance compared to the machine learning algorithm, Linear Regression.

REFERENCES

Abdul-Wahab, S. A., & Al-Alawi, S. M. (2002). Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. *Environmental Modelling & Software*, *17*(3), 219–228.

Coman, A., Ionescu, A., & Candau, Y. (2008). Hourly ozone prediction for a 24-h horizon using neural networks. *Environmental Modelling & Software*, *23*(12), 1407–1421.

Ettouney, R. S., Mjalli, F. S., Zaki, J. G., El-Rifai, M. A., & Ettouney, H. M. (2009). Forecasting of ozone pollution using artificial neural networks. *Management of Environmental Quality: An International Journal*, *20*(6), 668–683.

Jacoby, D., Ostrometzky, J., & Messer, H. (2021). Short-term prediction of the attenuation in a commercial microwave

link using LSTM-based RNN. *2020 28th European Signal Processing Conference (EUSIPCO)*, 1628–1632.

Kök, İ., Şimşek, M. U., & Özdemir, S. (2017). A deep learning model for air quality prediction in smart cities. *2017 IEEE International Conference on Big Data (Big Data)*, 1983–1990.

Lyu, P., Chen, N., Mao, S., & Li, M. (2020). LSTM based encoder-decoder for short-term predictions of gas concentration using multi-sensor fusion. *Process Safety and Environmental Protection*, *137*, 93–105.

Pires, J. C. M., Gonçalves, B., Azevedo, F. G., Carneiro, A. P., Rego, N., Assembleia, A. J. B., Lima, J. F. B., Silva, P. A., Alves, C., & Martins, F. G. (2012). Optimization of artificial neural network models through genetic algorithms for surface ozone concentration forecasting. *Environmental Science and Pollution Research*, *19*, 3228–3234.

Sadhukhan, S., & Yadav, V. K. (2023). Forecasting, capturing and activation of carbon-dioxide (CO $ _2 $): Integration of Time Series Analysis, Machine Learning, and Material Design. *ArXiv Preprint ArXiv:2307.14374*.

Zoran, M. A., Savastru, R. S., Savastru, D. M., & Tautan, M. N. (2020). Assessing the relationship between ground levels of ozone (O3) and nitrogen dioxide (NO2) with coronavirus (COVID-19) in Milan, Italy. *Science of The Total Environment*, *740*, 140005.