Journal of Applied Informatics and Computing (JAIC)
Vol.8, No.1, July 2024, pp. 22~29
e-ISSN: 2548-6861

Forecasting Air Quality Indeks Using Long Short Term Memory

Irfan Wahyu Ramadhani [1*], Filmada Ocky Saputra [2], Ricardus Anggi Pramunendar [3**], Galuh Wilujeng Saraswati [4**], Nurul Anisa Sri Winarsih [5**], Muhammad Syaifur Rohman [6**], Danny Oka Ratmana[7**], Guruh Fajar Shidik [8**]**

[*] Faculty of Computer Science, Dian Nuswantoro University, Semarang, 50131, Indonesia
[**] Research Center Of Intelligent Distributed Surveillance and Security, Dian Nuswantoro University, Semarang, 50131, Indonesia
iramadhani679@gmail.com [1], filmada.os@dsn.dinus.ac.id [2], ricardus.anggi@dsn.dinus.ac.id [3], galuhwilujengs@dsn.dinus.ac.id [4], nurulanisasw@dsn.dinus.ac.id [5], syaifur@dsn.dinus.ac.id [6], rdannyoka@dsn.dinus.ac.id [7], guruh.fajar@research.dinus.ac.id [8]

Article Info

Article history:

Received 2024-03-22
Revised 2024-05-08
Accepted 2024-05-15

Keyword:

Air Quality Index,
Air Pollution,
Forecasting,
LSTM,
Sports.

ABSTRACT

Exercise offers significant physical and mental health benefits. However, undetected air pollution can have a negative impact on individual health, especially lung health when doing physical activity in crowded sports venues. This study addresses the need for accurate air quality predictions in such environments. Using the Long Short-Term Memory (LSTM) method or what is known as high performance time series prediction, this research focuses on forecasting the Air Quality Index (AQI) around crowded sports venues and its supporting parameters such as ozone gas, carbon dioxide, etc. -others as internal factors, without involving external factors causing the increase in AQI. Preprocessing of the data involves removing zero values and calculating correlations with AQI and the final step performs calculations with the LSTM model. The LSTM model which adds tuning parameters, namely with epoch 100, learning rate with a value of 0.001, and batch size with a value of 64, consistently shows a reduction in losses. The best results from the AQI, PM2.5, and PM10 features based on performance are MSE with the smallest value of 6.045, RMSE with the smallest value of 4.283, and MAE with a value of 2.757.

I. INTRODUCTION

The impact of air pollution on air quality is a matter of global concern [1], [2]. In developing countries, especially in densely populated and industrialized areas, air pollution is on the rise, leading to a decline in air quality [3]. The repercussions of this decline result in both humans and the environment experiencing its negative effects [4], [5], [6], [7]. According to a report from the World Health Organization, as many as 7 million individuals are at risk of health threats due to the impact of air pollution [5], [6]. For example, air pollution along roads has an adverse impact on the lungs of individuals who exercise near roads [8]. and other examples also show the negative impact of certain gases such as ozone, which impacts lung function who exercise outdoors [9]. Individuals with lung or heart issues, children, and late teenagers are groups at a higher risk of negative impacts from air pollution [10].

Sport is one of the activities that makes the body healthy. Exercise has many positive impacts, thus triggering the popularity of engaging in physical activities both indoors and outdoors to maintain health [11]. Outdoor exercise has been shown in numerous studies to effectively lower depression and stress levels [11]. However, people often overlook the dangers of outdoor air pollution. Engaging in physical activity enhances bodily performance, especially the lungs that take in more air, thereby potentially increasing the risk of exposure to air pollution [12]. Certainly, this poses a danger to the body's health.

Long Short-Term Memory (LSTM) is one of the algorithms suitable for air quality forecasting. Hochreiter and Schmidhuber proposed long short-term memory (LSTM) in 1997 [13]. LSTM is a unique form of Recurrent Neural Network (RNN) equipped with internal memory and a multiplication gate [14]. LSTM-based models effectively address the challenge of vanishing gradients and the issues in learning long-term dependencies faced by traditional RNNs

[15]. Based on that, LSTM becomes a suitable algorithm for long-term time-series data.

Several studies related to forecasting for predicting air quality involving the LSTM algorithm have been conducted extensively. The study by Ekta Sharma et al.[16] on air quality forecasting for suspended particulate matter using Convolutional Neural Networks and LSTM resulted in performance with an average MAE of 6.4025, RMSE with an average of 20.535, and MAPE with an average of 30.4475, where the averages are taken from 4 locations.

Research by Yong-Chao Jin et al. [17] conducted a comparison of the ARIMA, LSTM, and improved hybrid ARIMA-LSTM algorithms in predicting COVID-19 data with performance metrics. In the prediction for Germany, the performance metrics were MSE with a value of 51566389.024, RMSE with a value of 36375.064, and MAE with a value of 17312.186. For the prediction in Japan, the performance metrics were MSE with a value of 20582526.517, RMSE with 4536.797, and MAE with a value of 2412.680.

Research by Baowei Wang et al. [18] conducted air quality forecasting based on the GRU and LSTM algorithms in IoT. Specifically, the LSTM algorithm exhibited training performance with RMSE ranging from 24 to 30 and testing performance with RMSE ranging from 18 to 20.

Research by Taoying Li et al.[19], which focused on forecasting particulate matter (PM2.5) using a combined CNN-LSTM model, revealed performance results for the LSTM algorithm. The multivariate LSTM alone achieved an average MAE of 15.324, while the univariate LSTM had an average MAE of 19.9454. The average RMSE performance for multivariate and univariate models was 18.0852 and 23.2646, respectively.

Research by Dinesh Komarasamy et al. [20], which focuses on the prediction and classification of air quality using machine learning and deep learning, achieved performance on Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN) with up to 98% accuracy.

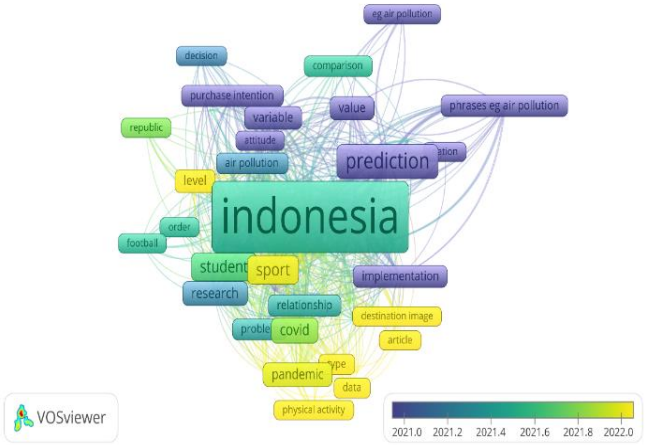

Figure 1. Visualization GAP Research

This research focuses on forecasting the air quality index at crowded sports venues in Pati, Central Java, using the Long Short-Term Memory (LSTM) algorithm. This research aims to provide air quality forecast information and assess the effectiveness of the LSTM algorithm in predicting the air quality index at crowded sports venues in Pati, Central Java. The air quality forecast information is intended to be one of the steps in mitigating the negative impacts of air pollution.

This research also helps add information regarding forecasting the air quality index in Indonesia, which is still relatively small based on VOS Viewer visualization. VOS viewer, as a statistical tool, assesses the trends and orientations within a specific domain at a particular stage [21]. This involves conducting analyses such as author-institution co-occurrence, keyword statistics, and visually mapping co-cited literature, which is then clustered and illustrated based on their respective weights [21]. Visualization of 200 journals on Google Scholar with the keywords "air quality prediction, sports, Indonesia". The visualization results are depicted in Figure 1.

II. RESEARCH METHOD

The entire research flow as a whole can be seen in Figure 2. The research begins with data collection and proceeds to the data preprocessing stage. The creation of the LSTM model will be tailored to the research, and the data resulting from preprocessing will be fed into the LSTM model for calculation. This process aims to obtain forecasting results and assess the performance of the created model.

A. Data Collecting

The data collection stage is depicted in Figure 2 in red. The dataset consists of publicly available data obtained from a website (https://www.weatherbit.io/). The dataset includes hourly air quality observations spanning approximately one year from January 12, 2022, to December 31, 2022, at four crowded sports locations in Pati, Central Java. The locations are in the City Square of Pati, around GOR (sports center) Pesantenan Pati, Joyokusumo Football Field, and Safin Academy Football Field. One of the fastest ways to collect data on internet website pages is by the crawling method [22], as used [23] to speed up research. This method works by collecting data and information automatically by filtering all internet pages to create an index of the information he is looking for. However, this research was carried out manually. The data retrieval technique is done manually by making requests to the API according to the instructions shown on the website. This is done to prevent the loss of some data that is suddenly lost when using the crawling method. Lastly, the data only uses data that includes internal factors causing the increase in AQI, so it does not use external factors such as weather conditions, wind direction, traffic conditions, and so on.

B. Data Preprocessing

The data preprocessing stage is indicated in Figure 2 in green. In data preprocessing itself, it consists of several steps. The raw data will be selected for the most influential features related to air pollution indicators. Key elements influencing air pollution encompass volatile organic compounds (VOCs), carbon oxides (COx), Sulfur oxides (SOx), Nitrogen oxides (NOx), and ozone (O3). These factors collectively represent the most impactful contributors to the degradation of air quality [24]. This research utilizes the selected features, namely CO, NO2, O3, SO2, PM10, and PM2.5, as well as additional features for forecasting, which are AQI and datetime.

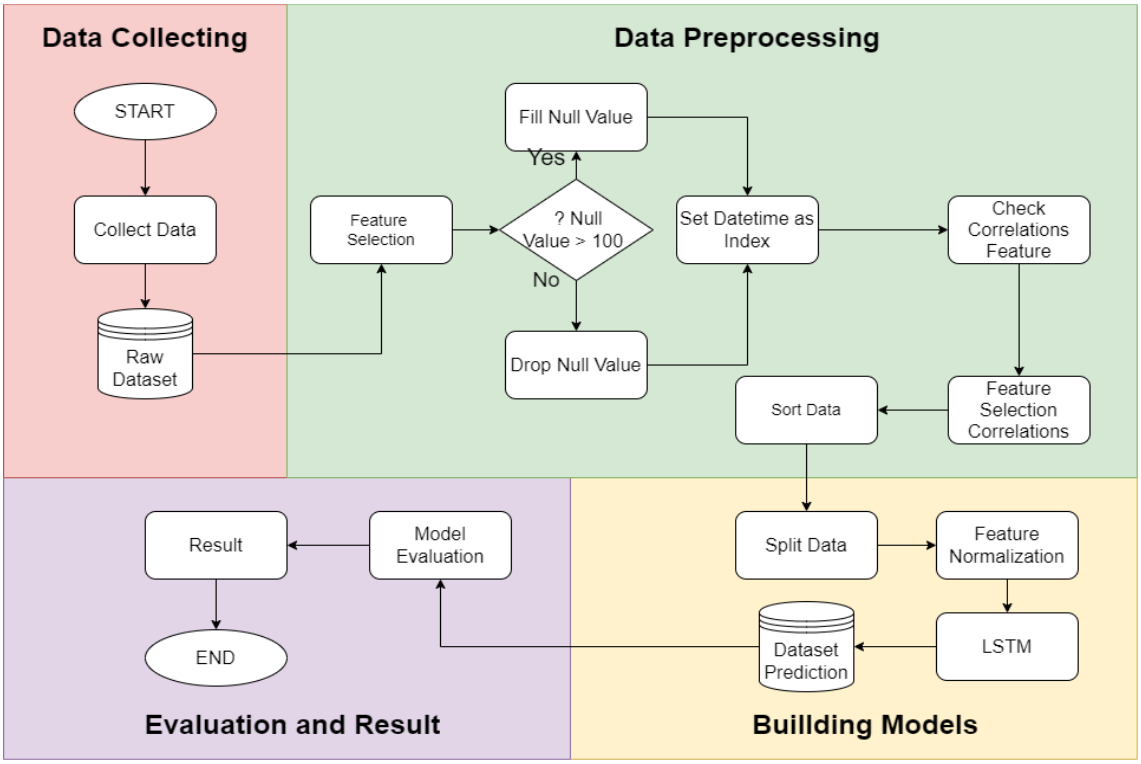

Figure 2. Workflow of Research Stages

The selected data will be examined for missing values. If there are indications of missing values, the next step is to fill in these missing values. If the observed missing values are less than 100, then the data with missing values will be deleted. If not, the features SO2, CO, O3, NO2, PM10, and PM2.5 will be filled with their respective median values to mitigate the impact of extreme values in the data. As for the AQI feature, it will be filled with the result of a mathematical calculation as indicated in Formula 1.

$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}} \times (C - C_{low}) + I_{low} \quad (1)$$

In the context provided, C stands for the pollution concentration, where C_{low} represents concentrations equal to or below a specific breakpoint, and C_{high} indicates concentrations equal to or exceeding another breakpoint. For I, it is the Air Quality Index (AQI) with I_{low} being the corresponding breakpoint index to C_{low}, and I_{high} being the corresponding breakpoint index to C_{high}. If each feature has obtained an air quality index, then the highest value will be used to fill the AQI feature formulated in Formula 2.

$$\max(x_1, x_2, x_3, \dots, x_i) \quad (2)$$

The next step is to change the datetime feature into an index to create time-series data. Finally, we calculate the correlation using the Pearson method from all features to the AQI feature. Pearson's method is used because it is the default function of the Python code. Pearson correlation is shown in Formula 3.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (3)$$

Where x_i with x as the sample variable, $\bar{x}$ is the mean of x, y_i with y as the sample variable, and $\bar{y}$ is the mean of y. Only correlation results above 0,8 with the AQI feature will be used in the LSTM model calculation because if the correlation is lower than that, the feature may have less impact on the AQI feature. After obtaining features with correlations above 0.8, the data will be further selected by retaining features with correlations above 0.8, as well as the AQI feature. The process is then followed by the final step, which is sorting the data based on the datetime/index order.

C. Long Short-Term Model

This stage is the building model stage. Which is shown in yellow in Figure 2. and the LSTM structure can be seen in Figure 3. Three steps in the operation of the LSTM algorithm [25]. First, passing through the forget gate, which decides which cell state to discard, either the hidden state or new data, can be formulated by the calculation in Formula 4. Second, passing through the input gate, which checks the previous data from the first step, and if it is deemed worthy of being stored in the cell state, the gate decides on the new information to be added.

It can be formulated with formula 6 for recognizing new information and formula 5 for incorporating it. Finally, at the output gate, deciding on the new hidden state is formulated with formula 8. Formula 9 represents the formulation for the new hidden state along with the new cell state after the final process. In Figure 3, σ is the sigmoid function found in point 10, which produces values between 0 and 1. Here, 0 means "stopping all flow" while 1 means "letting all flow through." Then, the hyperbolic tangent function described in Formula 11 is used to address the issue of vanishing gradients.

$$F(t) = \sigma(W_f \cdot [H_{t-1}, X_t] + bf) \quad (4)$$

$$I(t) = \sigma(W_i \cdot [H_{t-1}, X_t]) + bi) \quad (5)$$

$$\tilde{C}(t) = \tanh(W_c \cdot [H_{t-1}, X_t] + bc) \quad (6)$$

$$C(t) = ft * C_{t-1} + It * \tilde{C}t \quad (7)$$

$$O(t) = \sigma(Wo \cdot [H_{t-1}, X_t] + bo) \quad (8)$$

$$H(t) = Ot * \tanh(Ct) \quad (9)$$

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}} \quad (10)$$

$$\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad (11)$$

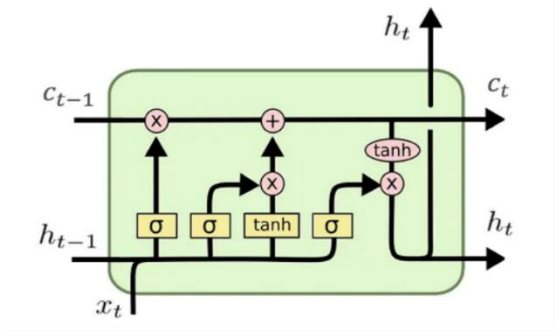

Figure 3. LSTM Structure

The input weights are represented by W_f, W_i, W_c, and W_o, and the bias weights are denoted by bf, bi, bc, and bo. The current time state is symbolized by t, while $t-1$ signifies the previous time. Input is represented by X, output by H, and the cell state by C.

In the stage of creating models or the yellow-colored stage in Figure 2, it consists of several steps. The data resulting from preprocessing is multivariate. During the data splitting phase, the dataset undergoes division into two segments: one designated for training and the other for testing. This division applies to all features that have undergone preprocessing.

The next stage is normalizing the split data. normalize data using Min-Max Normalization. Min-max normalization is an approach where the original data undergoes linear transformations to ensure a uniform scale, promoting equitable comparisons of values both before and after the normalization process [26]. Data normalized with Min-Max will be data in the range 0 to 1 using the formulation in formula 12. This normalization is intended to maintain model performance, and reduce the risk of missing gradients or exploding gradients.

$$X_{scaled} = \frac{X-X_{min}}{X_{max}-X_{min}} \quad (12)$$

Where X_{min} is the minimum number of X features, and X_{max} is the maximum number of X features.

The final stage is to create an LSTM model with layers that are used for air quality forecasting calculations. The model is sequential with the first layer being the LSTM layer, the second layer being the Repeat Vector layer, the third layer being LSTM, the fourth layer being the dropout layer, and the last layer being the Time Distributed layer.

D. Model Evaluation

Figures The model will be evaluated by the results of forecasting calculations and the performance of the algorithm including Loss, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Sequentially all the performances are formulated in points 13, 14, 15.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \quad (13)$$

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2} \quad (14)$$

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \tilde{y}_i| \quad (15)$$

Where n is the length of the data and y is the original data and $\bar{y}$ is the estimated data.

III. Result and Discussion

A. Data Collecting Results

Results of manual data collection from websites (https://www.weatherbit.io/) for a sample of four crowded sports locations yielded a total of 8196 data points for each location, totaling 32784 for all four locations. The time range

for the data collected is from 17:00 on January 12, 2022, to 17:00 on December 30, 2022. For the visualization of these four crowded sports venues, please refer to Figure 4. The location placement in the datasets is determined by their geographical coordinates, specifically longitude and latitude, providing a precise representation of their positions on the Earth's surface. With a red pin indicating the location of Pati city square, a blue pin for the surroundings of GOR (sports center) Pesantenan Pati, a green pin for joyokusumo field, and a purple pin for safin Academy.

B. *Preprocessing Result*

From four datasets, all of them have the same features, namely city_name, country_code, lat, lon, state_code, timezone, aqi, co, datetime, no2, o3, pm10, pm25, so2, timestamp_local, timestamp_utc, and ts. Then feature selection is carried out and the remaining features are left co, co2, o3, so2, pm10, and pm25 and the most important features are aqi and datetime. With this the data is clean from features that will not be used.

Features that have been selected and see indications of null values. From four datasets, all of them indicated null values for the aqi, co, no2, o3, pm10, pm25, and so2 features with 55 null values for each data. because the null value for each feature is below 100, the value indicated as null will be removed from each feature that has a null value. Null values were removed because the values were quite small compared to the total data, which was 8196 for each dataset.

The four datasets are not yet in the form of a timeseries, so the datetime feature is used as an index to make them into a timeseries. and the dataset that has been made into a timeseries looks at the correlation with the aqi feature and obtained features pm10, and pm25 are features that have a high correlation and almost similar values. for pm10 with an average correlation of 0.996 and pm25 with an average correlation of 0.993. We separate the pm10 and pm25 features along with the aqi by selecting features, and finally the data is sorted based on the datetime index. From the results of this preprocessing, PM10 and PM25 are the features that have the most influence on the air quality index in the four samples of popular sports venues.

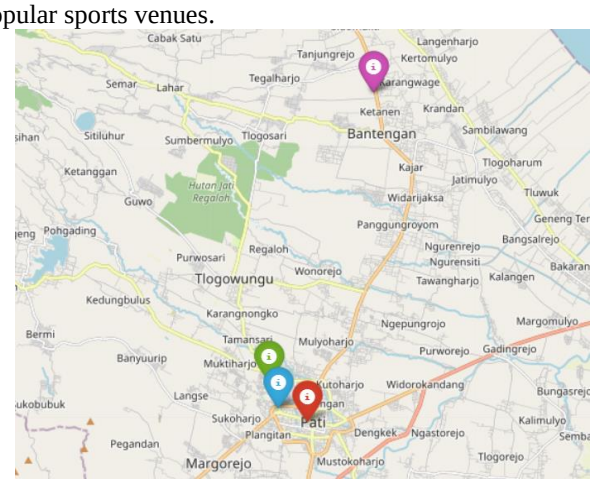

Figure 4. Location Mapping

C. *Model Result*

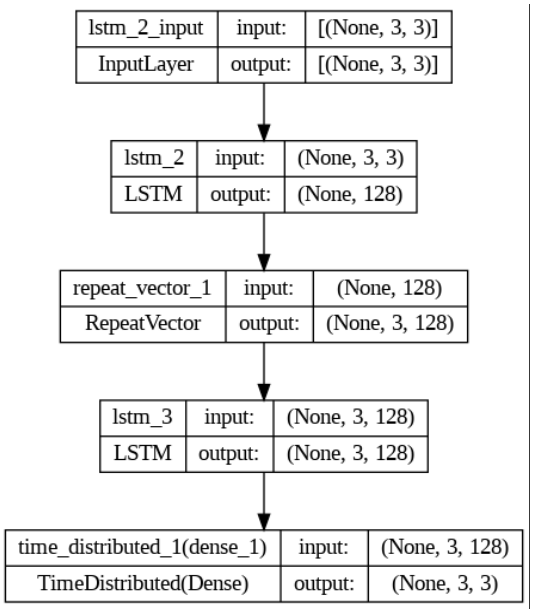

Figure 5. LSTM Models

The model with the LSTM algorithm is visualized in Figure 5. First, the data is input according to the shape size, namely 3*3 and continues to output with the same shape. input and output based on the number of features used. The data processed here will proceed to the first layer. On the first layer, LSTM calculations are carried out as in Figure 4 with input shape 3*3 and produces a shape with none (not specified) and filter 128. A popular activation function used in deep learning is ReLU [27]. In this layer ReLU activation is used to change negative values to 0 and positive values remain ReLU calculated using the formulation in formula 16.

$$f(x) = \max(0, x) \tag{16}$$

In this context, x refers to the input of the ReLU function, and $f(x)$ is the output. If x is positive, the output value will be equal to x, while if x is negative or zero, the output value will be zero. In the second layer, apply Repeat Vector to repeat the vector matrix as many times as the output output is none*3*128. filter 128 is obtained from the input to Repeat Vector. for the third layer the same as the first layer. The last layer applies time distributed with a deep number of features to create a sequence of data in processing related to time.

The model also uses tuning parameters as external parameters to improve model performance. The complete tuning parameters are shown in table 1. With Learning rate determining the number of steps per iteration, batch size is the number of samples per iteration. Activation uses ReLu to overcome the missing gradient or replace negative values with 0 and positive values remain. ReLu is formulated in formula 16. The loss function with Categorical Cross Entropy is used in multiclass training to calculate the model probability distribution with the actual data probability.

Categorical Cross Entropy is formulated in formula 17. and the optimizer uses Adam (Adaptive Moment Estimation) to optimize the model created.

$$L(y,p) = -\sum_{i=1}^{K} p_i \cdot \log(y_i) \quad (17)$$

TABLE I
HYPERPARAMETER VALUES

Hyperparameter	Values
Iteration	100 & 200
Learning rate	0.1, 0.01, & 0.001
Batch size	64
Activation	ReLu
Loss Function	Categorical cross Entropy
optimizer	Adam

D. Model Evaluation

Evaluation results based on the four datasets and the training process using the LSTM model. Results will include the performances of the dataset against the model, hyperparameters, and loss function visualization. Sequentially, the training results on the four datasets along with performance calculations can be seen in tables 2, 3, 4, 5. With the performance calculation results being the first numbers obtained after one process.

Each dataset stops at epoch 100. This is because if the epoch is more than that, the data becomes overfitting. Meanwhile, if you use epoch below. So, the table only shows epoch up to 100. The number of trains and tests here is just a comparison of the numbers, while still paying attention to the order of the data. This number is much better than dividing the number of trains and tests into 60:40, 80:20, or 90:10. This study does not use a 50:50 ratio, the results are very random rather than the number of comparisons already mentioned.

This research uses learning rates with values of 0.1, 0.01, and 0.001 and the results are very optimal. When using a learning rate of 0.1, the data train produces quite stable results and quite good performance. However, these results cannot be used directly, because they require comparison with other learning rate values. then a learning rate of 0.01 is used as a comparative learning rate value. When using a learning rate with a value of 0.01, performance results show a significant increase in performance.

To check whether the learning rate is stable, of course another learning rate value is needed as a third comparison value. The value 0.001 is used as a third learning rate comparison. The performance results with a learning rate of 0.001 are very good for use in the model. The performance of all error calculations shows that the results are mostly stable and have experienced a good decrease. except for the Joyokusumo field dataset which experienced a slight decrease in performance. The learning rate that has been mentioned is run on a model with a number of epochs of 100, because with epochs of 200, the pattern of decline and stability of some performance does not show this value and also epoch 200 is not displayed, so the best results for that time are with epoch 100 and learning rate 0.001 in the model calculation results.

Each dataset produces different performance. This performance is a calculation of the error from the prediction or forecasting results in the LSTM model, namely MSE, RMSE, and MAE. These performance values are the average values from each epoch, and use different learning rates. The error calculation result is the first result of a single calculation process on the model created. So that this value is likely to change according to the second, third, and so on processes of the model used. the change in value may not be very significant depending on the stability of the model.

From the MSE, you can see the difference in the square value between the original data and the predicted data. The calculation produces quite significant error values in the four datasets. The MSE value decreases as the learning rate of the three features decreases. except for the Joyokusumo field dataset with a learning rate of 0.001.

Calculating RMSE is the same as calculating MSE. From the four datasets, it can be seen that the RMSE values tend to be small. All features in the four datasets show stability in the RMSE calculations. From the MAE value in each dataset, it can be seen that the results are quite stable for each epoch and learning rate used. Mae values tend to decrease according to the learning rate used.

Based on performance calculation data, the best results are with a learning rate of 0.001, so the prediction results displayed are only based on this calculation. The dataset with the best prediction results and performance can be seen in Figure 8. With the first graph being the prediction of the AQI feature, the second graph being the PM10 feature, and the last being the PM2.5 feature. Also, value which represents the value of each data based on an index, where the index represents the dates in sequence.

TABLE II
PATI SQUARE DATASET

Epoch	Train	Test	Learning rate	PM10			PM2.5			AQI		
				MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE
100	70	30	0.1	294.613	17.164	13.453	141.22	11.883	9.312	711.409	26.672	21.152
100	70	30	0.01	42.624	6.528	4.673	19.047	4.364	3.062	96.582	9.827	6.759
100	70	30	0.001	36.55	6.045	3.985	18.806	4.336	2.882	94.583	9.725	6.298

TABLE III
PESANTENAN SPORTS CENTER DATASET

Epoch	Train	Test	Learning_rate	PM10			PM2.5			AQI		
				MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE
100	70	30	0.1	50.138	7.080	4.910	23.555	4.853	3.341	98.676	9.933	6.554
100	70	30	0.01	39.296	6.268	4.075	18.138	4.258	2.717	106.511	10.320	7.169
100	70	30	0.001	39.643	6.304	4.139	18.351	4.283	2.757	99.643	9.982	6.807

TABLE IV
JOYOKUSUMO FIELD DATASET

Epoch	Train	Test	Learning_rate	PM10			PM2.5			AQI		
				MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE
100	70	30	0.1	42.297	6.503	4.222	84.893	9.213	6.013	212.826	14.588	10.030
100	70	30	0.01	42.470	6.516	4.421	18.824	4.338	2.833	94.248	9.708	6.036
100	70	30	0.001	44.433	6.665	4.528	21.355	4.621	3.109	118.943	10.906	7.605

TABLE V
SAFIN ACADEMY DATASET

Epoch	Train	Test	Learning_rate	PM10			PM2.5			AQI		
				MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE
100	70	30	0.1	76.242	8.731	5.984	36.325	6.027	4.081	173.153	13.158	9.329
100	70	30	0.01	41.479	6.440	4.303	19.076	4.367	2.858	123.466	11.111	7.851
100	70	30	0.001	41.346	6.430	4.337	18.884	4.345	2.866	112.931	10.626	7.452

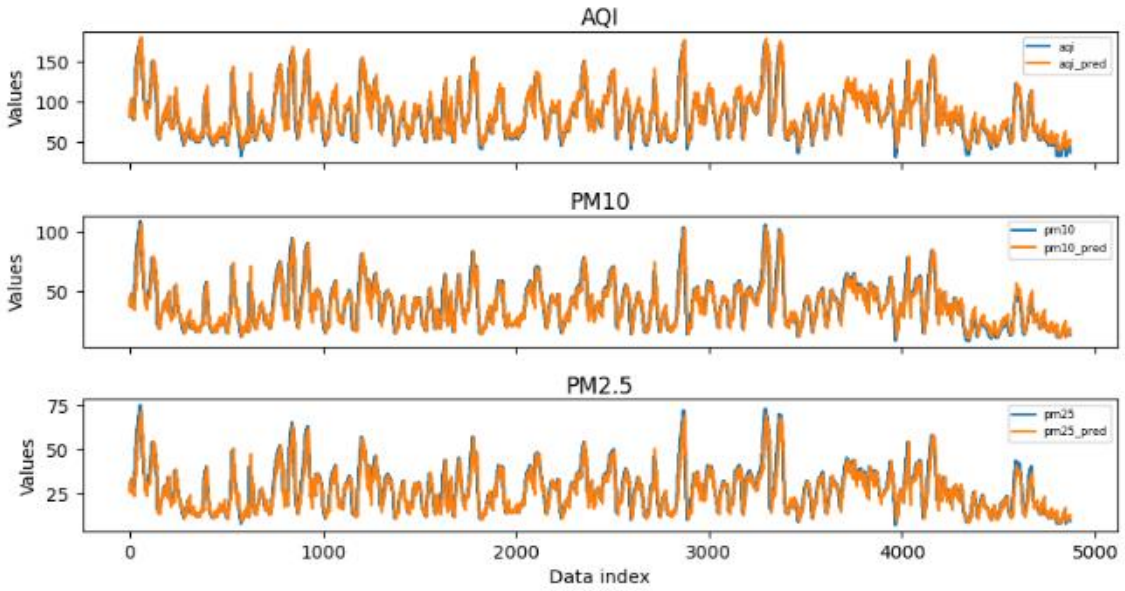

Figure 5. Result Model with Best Forecasting

IV. CONCLUSION

Research into air quality forecasting has yielded important insights, with a focus on model performance and forecast results, with the LSTM algorithm emerging as a standout performer. In particular, on four datasets featuring AQI features, the LSTM algorithm shows commendable results, with a minimum Mean Squared Error (MSE) of 94,248, Root Mean Squared Error (RMSE) at an impressive 9,708, and Mean Absolute Error (MAE) at a noteworthy 6,036. These results, obtained from careful trial-and-error experiments, underline the stability and reliability of the algorithm.

Exploration of various hyperparameters, including changes in train test splits ranging between 60 - 40, 70 - 30, 80 - 20, and 90 - 10, revealed that the most optimal and consistent results were achieved with a 70-30 trial split. Significantly, these results surpass the performance of several studies referenced in the introduction, underscoring the power and flexibility of LSTM algorithms in air quality estimation.

By examining the forecast results, the LSTM model shows commendable predictive capabilities, especially in projecting

the air quality index at popular sports venues in Pati, Central Java. This promising performance indicates the suitability of the LSTM algorithm for accurate and reliable air quality estimation, thereby highlighting its potential significance in improving environmental monitoring efforts. especially in the Pati area.

ACKNOWLEDGEMENT

Thank you to Dian Nuswantoro University as the author's home university. Thank you also to all the lecturers at Udinus. and also, thanks to IDSS Udinus as the person who supervised the author in writing this study.

REFERENCES

[1] H. Chen, M. Guan, and H. Li, "Air Quality Prediction Based on Integrated Dual LSTM Model," *IEEE Access*, vol. 9, pp. 93285–93297, 2021, doi: 10.1109/ACCESS.2021.3093430.

[2] S. M. Grath, E. Garrigan, and L. Zeng, "Predicting Air Quality Index Using Deep Neural Networks," in *2021 IEEE International Conference on Electronic Technology, Communication and Information, ICETCI 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 341–344. doi: 10.1109/ICETCI53161.2021.9563356.

[3] R. Saha, S. N. M. A. Hoque, M. M. R. Manu, and A. Hoque, "Monitoring Air Quality of Dhaka using IoT: Effects of COVID-19," in *International Conference on Robotics, Electrical and Signal Processing Techniques*, 2021, pp. 715–721. doi: 10.1109/ICREST51555.2021.9331026.

[4] A. Catovic, E. Kadusic, C. Ruland, N. Zivic, and N. Hadzajlic, "Air pollution prediction and warning system using IoT and machine learning," in *International Conference on Electrical, Computer, Communications and Mechatronics Engineering, ICECCME 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICECCME55909.2022.9987957.

[5] A. Desai, E. Gujarathi, S. Parikh, S. Yadav, Z. Patel, and N. Batra, "Deep Gaussian Processes for Air Quality Inference," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jan. 2023, pp. 278–279. doi: 10.1145/3570991.3571004.

[6] S. Ameer *et al.*, "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities," *IEEE Access*, vol. 7, pp. 128325–128338, 2019, doi: 10.1109/ACCESS.2019.2925082.

[7] Y. Wang and T. Kong, "Air Quality Predictive Modeling Based on an Improved Decision Tree in a Weather-Smart Grid," *IEEE Access*, vol. 7, pp. 172892–172901, 2019, doi: 10.1109/ACCESS.2019.2956599.

[8] N. Syed *et al.*, "Effects of Traffic-Related Air Pollution on Exercise Endurance, Dyspnea, and Cardiorespiratory Responses in Health and COPD," *Chest*, vol. 161, no. 3, pp. 662–675, Mar. 2022, doi: 10.1016/j.chest.2021.10.020.

[9] A. Hung, H. Nelson, and M. S. Koehle, "The Acute Effects of Exercising in Air Pollution: A Systematic Review of Randomized Controlled Trials," *Sports Medicine*, vol. 52, no. 1, pp. 139–164, Jan. 2022, doi: 10.1007/s40279-021-01544-4.

[10] E. Brumancia, S. Justin Samuel, and L.Mary Gladence, "Air Pollution Detection and Prediction Using Multi Sensor Data Fusion," in *Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020) : 13-15 May, 2020*, Madurai, India: IEEE, Jun. 2020, pp. 844–849.

[11] F. Qin *et al.*, "Exercise and air pollutants exposure: A systematic review and meta-analysis," *Life Sciences*, vol. 218. Elsevier Inc., pp. 153–164, Feb. 01, 2019. doi: 10.1016/j.lfs.2018.12.036.

[12] B. Marmett, R. B. Carvalho, G. P. Dorneles, R. B. Nunes, and C. R. Rhoden, "Should I stay or should I go: Can air pollution reduce the health benefits of physical exercise?," *Med Hypotheses*, vol. 144, Nov. 2020, doi: 10.1016/j.mehy.2020.109993.

[13] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural Computation*, vol. 31, no. 7. MIT Press Journals, pp. 1235–1270, Jul. 01, 2019. doi: 10.1162/neco_a_01199.

[14] K. Smagulova and A. P. James, "A survey on LSTM memristive neural network architectures and applications," *European Physical Journal: Special Topics*, vol. 228, no. 10. Springer Verlag, pp. 2313–2324, Oct. 01, 2019. doi: 10.1140/epjst/e2019-900046-x.

[15] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The Performance of LSTM and BiLSTM in Forecasting Time Series," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 3285–3292. doi: 10.1109/BigData47090.2019.9005997.

[16] E. Sharma, R. C. Deo, R. Prasad, A. V. Parisi, and N. Raj, "Deep Air Quality Forecasts: Suspended Particulate Matter Modeling with Convolutional Neural and Long Short-Term Memory Networks," *IEEE Access*, vol. 8, pp. 209503–209516, 2020, doi: 10.1109/ACCESS.2020.3039002.

[17] Y. C. Jin, Q. Cao, K. N. Wang, Y. Zhou, Y. P. Cao, and X. Y. Wang, "Prediction of COVID-19 Data Using Improved ARIMA-LSTM Hybrid Forecast Models," *IEEE Access*, vol. 11, pp. 67956–67967, 2023, doi: 10.1109/ACCESS.2023.3291999.

[18] B. Wang, W. Kong, H. Guan, and N. N. Xiong, "Air Quality Forecasting Based on Gated Recurrent Long Short Term Memory Model in Internet of Things," *IEEE Access*, vol. 7, pp. 69524–69534, 2019, doi: 10.1109/ACCESS.2019.2917277.

[19] T. Li, M. Hua, and X. Wu, "A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5)," *IEEE Access*, vol. 8, pp. 26933–26940, 2020, doi: 10.1109/ACCESS.2020.2971348.

[20] D. Komarasamy, G. C. Nandhidha, S. Nandhinidevi, P. L. Nanthini, Mohanasaranya, and K. Kousalya, "Air Quality Prediction and Classification using Machine Learning," in *Proceedings - 7th International Conference on Computing Methodologies and Communication, ICCMC 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 187–191. doi: 10.1109/ICCMC56507.2023.10083760.

[21] W. Xin and W. Bin, "Visual Analysis of Precision Teaching Research Based on VOS Viewer in Data-Driven Perspective," in *2023 IEEE 12th International Conference on Educational and Information Technology, ICEIT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 179–185. doi: 10.1109/ICEIT57125.2023.10107845.

[22] M. S. Rohman, H. A. Santoso, G. W. Saraswati, N. Anisa, and S. Winarsih, "Pemanfaatan Topic-Focused Crawler untuk Pembangunan Corpus Berita Bencana menggunakan Teknik Scrapy CSS Selector." [Online]. Available: http://bpbdjateng.com

[23] M. Alysha Zulia Larasati, N. Anisa Sri Winarsih, M. Syaifur Rohman, and G. Wilujeng Saraswati, "Penerapan Metode K-Means Clustering dalam Menganalisis Sentimen Masyarakat terhadap K-Popers pada Twitter," *Progresif: Jurnal Ilmiah Komputer*, vol. 18, no. 2, pp. 201–210, 2022, Accessed: Jan. 15, 2024. [Online]. Available: doi:10.35889/progresif.v18i2.877

[24] S. Masmoudi, H. Elghazel, D. Taieb, O. Yazar, and A. Kallel, "A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection," *Science of the Total Environment*, vol. 715, May 2020, doi: 10.1016/j.scitotenv.2020.136991.

[25] F. Hamami and I. A. Dahlan, "Univariate Time Series Data Forecasting of Air Pollution using LSTM Neural Network," in *2020 International Conference on Advancement in Data Science, E-Learning and Information Systems, ICADEIS 2020*, Institute of Electrical and Electronics Engineers Inc., Oct. 2020. doi: 10.1109/ICADEIS49811.2020.9277393.

[26] H. 1, T. Wahyuningsih , and E. Rahwanto, "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer." [Online]. Available: http://archive.ics.uci.edu/ml.

[27] G. F. Shidik *et al.*, "LUTanh Activation Function to Optimize BI-LSTM in Earthquake Forecasting," *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 1, pp. 572–583, Feb. 2024, doi: 10.22266/ijies2024.0229.48.