IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 3, June 2026, pp. 2543~2553
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i3.pp2543-2553

Machine learning approaches for anomaly detection of Jakarta air quality index

Muhammad Rizky Nurhambali, Yenni Angraini, Anwar Fitrianto
Statistics and Data Science Study Program, School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia

Article Info

Article history:

Received Feb 12, 2025
Revised Feb 9, 2026
Accepted May 12, 2026

Keywords:

Air quality index
Anomalies
Machine learning
Outliers
Time series

ABSTRACT

Anomalies in time series data are observations that deviate markedly from surrounding values or overall patterns. Air quality index (AQI) data, which vary over time, provide a suitable context for anomaly detection. Time series anomaly detection can be done with machine learning approaches like long short-term memory (LSTM) and extreme gradient boosting (XGBoost). These methods have advantages over conventional methods in handling nonlinearity and large data dimensions. This study compares LSTM and XGBoost methods for detecting anomalies in Jakarta's hourly AQI data. The dataset was obtained from the AirNow website and covers the period from January 1, 2018, to December 31, 2023. Anomalies in the observed data were labeled using moving range (MR) (2) and (3) approaches with three and four-sigma thresholds, and feature engineering (FE) was applied to improve model performance. The results indicate that LSTM is more suitable than XGBoost for forecasting and classification tasks in AQI data. LSTM achieved an average mean absolute percentage error (MAPE) of 10.3840%, a root mean square error (RMSE) of 10.5913, and a balanced accuracy (BACC) of 0.9424 under MR (2) labeling with the four-sigma rule. The anomalies detected mostly occurred between 21:00 and 09:00 and during the rainy season.

Corresponding Author:

Muhammad Rizky Nurhambali
Statistics and Data Science Study Program, School of Data Science, Mathematics, and Informatics
IPB University
Bogor, Indonesia
Email: rizkynurhambali@apps.ipb.ac.id

1. INTRODUCTION

An anomaly is a condition where a system does not work as usual and is significantly different from the general system [1]. Related to data, anomalies are commonly referred to as outliers. In time series data, an anomaly is defined as the presence of unexpected behavior in a time series data within a certain time interval [2]. However, in its application, anomalies in time series data are difficult to detect even though anomaly detection is an important thing to do considering that anomaly detection can be an indicator of future problems.

Anomaly detection in time series data evolved from the classical or conventional approach, which uses the difference between forecasting results and actual data. However, the weaknesses found have led to the development of various methods, including machine learning. Extreme gradient boosting (XGBoost) as an example of machine learning can handle non-linearity of time series with strong learning capabilities [3]. Furthermore, long short-term memory (LSTM) as one part of deep learning can ignore the stationarity assumption because it has the ability to handle nonlinear relationships and large dimensions [4]. In addition,

the specific model required by conventional methods in making predictions is not needed in the use of LSTM and other machine learning methods so that they can adapt better.

One example of an anomaly that may occur is air quality. Air quality is commonly represented by an air quality index (AQI) that reflects the level of air pollution. Air pollution has become a major environmental issue that is receiving increasing public attention, particularly alongside concerns related to climate change. Ghosh *et al*. [5] estimate that 5.8 million premature births in 2019 were caused by air pollution. Moreover, air pollution is known to cause various respiratory diseases, including pneumonia, asthma, and acute respiratory infections (ARI). Data indicate that between 2020 and 2022 there were 73,694 cases of pneumonia in children and 15,825 cases of asthma in children in Jakarta [6]. In addition, in the first semester of 2023 alone, 638,291 cases of ARI were reported in Jakarta [7]. Alla and Adari [8] noted that the AQI measured by multiple sensors across different locations and collected periodically through a centralized system inherently form time series data. Such time series data can be utilized as input for neural network–based models to perform anomaly detection using a time series-based approach.

There are many approaches to detecting anomalies in time series data, one of which is machine learning, which overcomes conventional methods' limitations. Wang *et al*. [9] used the LSTM method, which was compared with the K-means method on electric power consumption data, resulting in the LSTM method being a better method in terms of recall. Furthermore, in Liu's [10] research, anomaly detection was carried out on water distribution systems with various ensemble methods. The study showed the stability of XGBoost even though it was not the best method. Both methods have previously been compared by Trizoglou *et al*. [11] using wind turbine data. However, the use of both methods is still limited in detecting air quality anomalies, especially in labeled time series data, so the application and comparison of methods under these conditions is a novelty in this study. Therefore, this study is based on several questions from previous studies, namely how LSTM and XGBoost perform in detecting anomalies in Jakarta's labeled AQI, and how the occurrence of air quality anomalies is related to meteorological factors. Through this study, this research aims to provide insight into anomaly patterns in air quality data and evaluate the feasibility of machine learning methods in supporting air quality monitoring and early warning systems.

2. METHOD

2.1. Long short-term memory

LSTM is a type of neural network developed of recurrent neural network (RNN) by Hochreiter and Schmidhuber [12]. LSTM has internal memory (c) and multiplicative gates, namely the forget gate (f), input gate (i), and output gate (o), to overcome vanishing or exploding gradients in updating weights in RNN. This also gives LSTM the ability to process sequential data and store information both in the short and long term [13]. Therefore, LSTMs are widely used to process text and time series data [14].

Each multiplicative gate of the LSTM has a different function. The forget gate sorts out the information to be stored or deleted, the input gate updates the information in the internal memory, and the output gate regulates the output of information. Mathematically, the whole process in LSTM works following (1) to (6), where W is the weight, h is the hidden state, b is the bias, g is the activation function (commonly sigmoid and tanh), and the '$\odot$' operator indicates the multiplication of two vectors in the same direction [15].

$$f_t = g\left(W_{fh}h_{t-1} + W_{fx}x_t + b_f\right) \quad (1)$$

$$i_t = g(W_{ih}h_{t-1} + W_{ix}x_t + b_i) \quad (2)$$

$$\tilde{c}_t = g(W_{\tilde{c}h}h_{t-1} + W_{\tilde{c}x}x_t + b_{\tilde{c}}) \quad (3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (4)$$

$$o_t = g(W_{oh}h_{t-1} + W_{ox}x_t + b_o) \quad (5)$$

$$h_t = o_t \odot g(c_t) \quad (6)$$

2.2. Extreme gradient boosting

XGBoost is a machine learning algorithm developed by Chen and Guestrin [16]. It uses the gradient boosting method with an ensemble algorithm, which combines several algorithms (learners) to improve accuracy designed for computational efficiency and model flexibility. This method optimizes gradient boosting by creating a sequential decision tree that minimizes error. At each iteration, the error value will be

calculated to correct the basic algorithm in the previous iteration [17], [18]. The prediction result in XGBoost is obtained based on the summation of all results as shown in (7), where F is the space of regression trees, $f_k(x_i)$ is the result of k trees, and $\hat{y}_i$ is the prediction value of the i-th example for x_i. The value of $\hat{y}_i$ is then obtained based on (7).

$$\hat{y}_i = \sum_{k=1}^{n} f_k(x_i), \; f_k \in F \tag{7}$$

2.3. Metrics evaluation

Evaluation metrics are measures used to determine the goodness of a model, and forecasting and classification are no exception. Standard evaluation metrics used in forecasting are root mean square error (RMSE) and mean absolute percentage error (MAPE). Lower RMSE and MAPE values reflect higher forecasting accuracy and indicate reliable predicted values. RMSE and MAPE are calculated based on (8) and (9) where y_t is t-time value, $\hat{y}_t$ is t-time forecast value, and n is number of forecasting periods. Furthermore, the evaluation metric used for calculating the accuracy of classification results is balanced accuracy (BACC). According to Bej *et al*. [19], BACC is used because of the imbalance of data between anomaly and non-anomaly classes, so its use will be more meaningful than accuracy. The BACC value is calculated using (10) where true positive (TP) is a positive class that is correctly classified as a positive class, true negative (TN) is a negative class that is correctly classified as a negative class, false positive (FP) is a negative class that is classified as a positive class, and false negative (FN) is a positive class that is classified as a negative class.

$$RMSE = \sum_{t=1}^{n} \left(\frac{(y_t - \hat{y}_t)^2}{n} \right)^{\frac{1}{2}} \tag{8}$$

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right| \times 100\% \tag{9}$$

$$BACC = \frac{1}{2} \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP} \right) \tag{10}$$

2.4. Data

The AQI in Indonesia is measured using two approaches, namely the Indonesian and U.S. AQI. However, according to Pramana *et al*. [20], the U.S. AQI has been established globally and is aligned with the World Health Organization (WHO) air quality guidelines. Therefore, this study will refer to the use of the U.S. AQI, which has six levels: good (0-50), moderate (51-100), unhealthy for sensitive groups (101-150), unhealthy (151-200), very unhealthy (201-300), and hazardous (>300).

The data used are sourced from AirNow (www.airnow.gov) [21] and version 2.3.6 of the Prediction of Worldwide Energy Resources (POWER) project per hour on 2024/06/16 (https://power.larc.nasa.gov/data-access-viewer/) [22]. Data sourced from AirNow is the result of PM2.5 measurements at the U.S. Embassy Air Quality Monitoring Station in Central Jakarta as response variables. Furthermore, data sourced from NASA POWER is meteorological data in the form of temperature, rainfall, humidity, wind speed, wind direction, and surface pressure as explanatory variables. These variables are more clearly shown in Table 1. Each data used has a time unit in the form of hours using Western Indonesian time (UTC+7) with a time range of January 1, 2018, at 00:00 until December 31, 2023, at 23:00, and is downloaded manually on the available site.

Table 1. List of variables used

Types of meteorological variables	Variable (code)	Unit
	AQI	
Humidity/precipitation	Specific humidity at 2 meters (QV2M)	g/kg
	Relative humidity at 2 meters (RH2M)	%
	Precipitation corrected (PRECTOTCORR)	mm/hour
Wind/pressure	Wind speed at 10 meters (WS10M)	m/s
	Wind speed at 50 meters (WS50M)	m/s
	Wind direction at 10 meters (WD10M)	Degrees
	Wind direction at 50 meters (WD50M)	Degrees
	Surface pressure (PS)	kPa
Temperature	Dew/frost point at 2 meters (T2MDEW)	°C
	Temperature at 2 meters (T2M)	°C
	Wet bulb temperature at 2 meters (T2MWET)	°C

2.5. Data analysis procedures

The R and Python softwares were used in data analysis according to the flow in Figure 1. The analysis begins with data imputation because the AQI data obtained has 2,281 blank values. The imputation method used is adjusted to the amount of sequential missing data with the criteria: i) if missing value at one time, using linear interpolation; ii) if missing value ≤4 consecutive hours, using seasonal decomposition or seasonal splitting; and iii) if missing value >4 consecutive hours, using disaggregation or LSTM. The complete data was then explored using time series plots and correlation plots to obtain data characteristics and patterns. Then, the unlabeled AQI data is labeled with moving ranges of length 2 (MR (2)) and 3 (MR (3)), which are combined with the 3-sigma and 4-sigma rules. After labeling the data, two types of classes will be obtained: anomalies and non-anomalies. The anomaly values were replaced with the average of two nearby data, such as performing linear interpolation to fill in missing values. Next, the data is split into annual data and used for modeling.

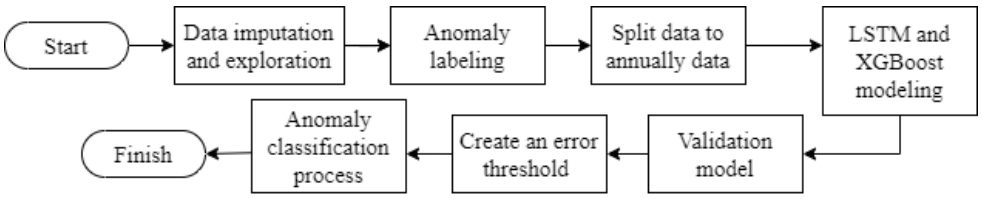

Figure 1. Flowchart of data analysis procedure

LSTM architecture is formed with two main layers and a dense layer. The hyperparameter values used in LSTM were determined by batch size (72), epoch (50), optimizer (Adam), and learning rate (0.001). The batch size and epoch values are determined subjectively by considering computation time. However, the optimizer and learning rate values are determined using sliding window and expanding window cross-validation. In the same way, the best XGBoost hyperparameters used were previously searched using sliding window and expanding window cross-validation, but no results were found that optimized model performance. Therefore, hyperparameters were used in Attaallah and Khan [23] research that modeled air quality with various machine learning approaches.

Validation is performed for each annual dataset using the LSTM and XGBoost models. The error (difference between the actual and forecast values) was calculated to obtain the method's accuracy and create an error threshold. Classification is performed on all the data based on the threshold obtained. The data will be classified as an anomaly if it exceeds the threshold.

3. RESULTS AND DISCUSSION

3.1. Data exploration

The complete AQI data was then explored. Descriptive statistics show that the average AQI for 2018-2023 is 98.15, which is categorized as moderate with a spread of 39.1388 from the mean. The Jakarta AQI reached the highest value of 219 on February 14, 2019, at 04:00, which indicates unhealthy condition in Jakarta. Meanwhile, the lowest AQI value occurred on March 27, 2018, at 02:00 with a value of 1 (good), indicating cleaner and healthier air.

Figure 2 shows that the AQI has a similar pattern every year. The AQI value tends to reach its peak in the morning around 09:00 and lowest in the evening around 19:00. The high morning AQI may be due to the high mobility of people in the morning who use vehicles to work or go to school. After the peak phase, the AQI decreases periodically and is high at night. According to Zheng *et al.* [24], there is a relationship between emissions and traffic in the variation of AQI where AQI at night is influenced by the increase in emissions from earth heating, vehicles, and corresponding particles accumulation.

Figure 3 shows the coefficient of correlation values between two variables using the “Spearman” method. The AQI has a negative correlation with almost all variables, except PS and T2M height. In addition, there are related explanatory variables, namely rainfall (PRECTOTCORR), humidity (QV2M and RH2M), dew point (T2MDEW), and wind direction (WD10M and WD50M). These results are in accordance with the research by Tian *et al*. [25] and Handhayani's [26], which discusses the relationship between meteorological factors and air quality. Even further, Handhayani's [26] mentioned that the relationship between rainfall and air quality can be utilized to make artificial rain to reduce the AQI.

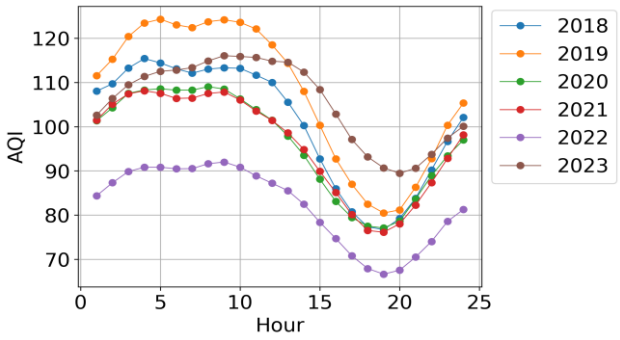

Figure 2. Average AQI for 24 hours per year

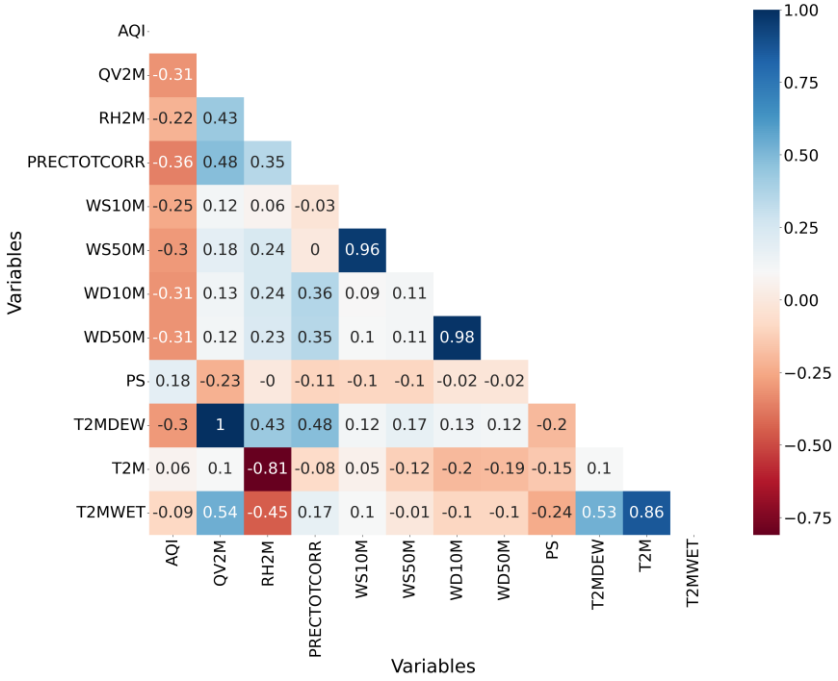

Figure 3. Correlation coefficient plot of variables

3.2. Validation model

Figure 4 shows the validation accuracy of LSTM and XGBoost based on the corresponding MAPE and RMSE values. The average MAPE and RMSE for LSTM validation results are 10.3840% and 10.5913. It is better than XGBoost, which has an average MAPE of 23.4267% and a RMSE of 21.3082. The low MAPE and RMSE values indicate that LSTM is able and better at capturing patterns in the data than XGBoost. This condition may also indicate that in the case of the AQI, XGBoost is unable to capture non-linear patterns as in the research of Rahman *et al*. [27] on data on COVID-19 cases in Bangladesh which shows XGBoost is no better than the autoregressive integrated moving average (ARIMA) method. In addition, based on research by Lv *et al*. [28], XGBoost is more stable for stationary data without outliers.

The addition of new explanatory variables was tried to improve the accuracy of XGBoost. The new explanatory variables are time variables (hour, day, and month) and numerical descriptive statistics, such as sum, mean, and quartiles of response variables using a rolling window for 24 hours. The results show an improvement in the MAPE and RMSE values of XGBoost, despite a decrease in LSTM. The decrease in the accuracy of the LSTM could be due to the loss of some information due to the use of a rolling window for the descriptive statistics of the AQI. In addition, the decrease in MAPE and RMSE values of LSTM may be an indication of anomalies in the data.

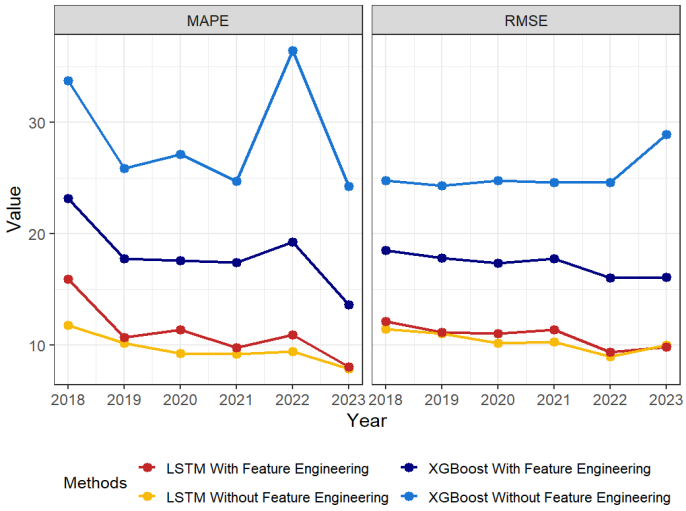

Figure 4. Validation accuracy of LSTM and XGBoost

3.3. Anomaly detection

3.3.1. Influences of moving range

In anomaly detection, MR plays more of a role in the actual data labeling process. The AQI does not have a definite label, so the MR of length 2 (MR (2)) and 3 (MR (3)) approaches are used. MR (2) clearly captures two consecutive observations as minimum and maximum, but MR (3) is not necessarily obtained from two consecutive observations. Furthermore, according to Woodall and Montgomery [29], using MR of length more than 2 will increase bias. Therefore, labeling the actual data with MR (2) can capture high intertemporal value changes compared to MR (3).

Table 2 generally shows that labeling with MR (2) is more suitable than MR (3) for the LSTM method. However, for the XGBoost method (as shown in Table 3), the MR (3) labeling method looks slightly better than MR (2) on average. This is indicated by the higher classification accuracy. The average classification accuracy of LSTM with MR (2) is 0.9924, while with MR (3) is 0.6933. Then, with XGBoost, the average classification accuracy with MR (2) is 0.5101 and MR (3) is 0.6349. Although there is a classification accuracy value of 1.00, it is due to the absence of observations labeled as anomaly classes in the actual data and the detection results also show similar results. Based on these results, LSTM outperforms XGBoost to capture fluctuating data.

3.3.2. Influences of sigma rules

The sigma rule in anomaly detection will affect the anomaly limit range as a threshold. As a result, the number of anomalies in the data also changes. The smaller the sigma value used, the narrower the non-anomaly region will be, causing a lot of data to fall outside the interval and vice versa. However, not only that, but the use of the sigma rule can also affect the classification accuracy in anomaly detection. Tables 2 and 3 shows the classification accuracy of the LSTM and XGBoost method is better using the 4-sigma rule. The average classification accuracy of LSTM with 3 and 4-sigma is 0.7485 and 0.8672 respectively, while XGBoost has classification accuracy values with 3 and 4-sigma of 0.5096 and 0.6408 respectively. These results are still in line with the research of Zheng *et al*. [30], where the use of 3 and 4-sigma is the best value among other values (1, 2, 5, 6, and 7).

Table 2. Classification evaluation for LSTM methods

Year	MR (2)				MR (3)			
	3-Sigma		4-Sigma		3-Sigma		4-Sigma	
	FE	WFE	FE	WFE	FE	WFE	FE	WFE
2018	0.9097	0.9313	0.9303	0.9718	0.5441	0.5916	0.9993	0.9997
2019	0.9292	0.9426	0.8617	0.8867	0.6004	0.5663	0.9995	0.9993
2020	0.8828	0.9275	0.9045	0.9641	0.5612	0.5776	0.6662	0.9997
2021	0.8633	0.9499	0.9302	0.9720	0.5245	0.5391	0.4996	0.4995
2022	0.8800	0.9178	0.9440	0.9454	0.6290	0.6432	0.4989	0.4932
2023	0.9207	0.9205	0.9386	0.9141	0.6008	0.6098	0.9978	0.9977

Table 3. Classification evaluation for XGBoost methods

Year	MR (2)				MR (3)			
	3-Sigma		4-Sigma		3-Sigma		4-Sigma	
	FE	WFE	FE	WFE	FE	WFE	FE	WFE
2018	0.5145	0.5023	0.4995	0.4999	0.4980	0.4982	0.9995	0.9998
2019	0.5218	0.5023	0.4997	0.4999	0.4973	0.4982	0.9994	0.9998
2020	0.5275	0.5054	0.5357	0.5117	0.5312	0.5139	0.6664	0.4998
2021	0.5059	0.5016	0.5000	0.5000	0.4974	0.4979	0.5000	0.5000
2022	0.5452	0.5088	0.5274	0.4999	0.5294	0.4967	0.4996	0.4999
2023	0.5195	0.5018	0.5122	0.5000	0.5176	0.4987	0.9999	1.0000

3.3.3. Influences of feature engineering

Feature engineering (FE) influences both methods in validation results, so its influence on classification accuracy is also studied. Tables 2 and 3 shows the complete classification results of the methods. When averaged, the LSTM method with the addition of FE has an accuracy of 0.7923, but without the addition of feature engineering (WFE) has an accuracy of 0.8234. Then, with FE, the average value of XGboost classification accuracy is 0.5810, while WFE, it is 0.5640. Both conditions show results that align with the validation process where adding FE makes the accuracy of the LSTM method decrease and XGBoost increase.

3.3.4. Air quality index anomaly detection results

The AQI has categories based on the PM2.5 indicator value. Based on Figure 5, anomaly detection with LSTM (Figure 5(a)) shows that many AQI anomalies occur in the unhealthy for sensitive group category. However, XGBoost captures many anomalies in the good and unhealthy categories as shown in Figure 5(b). As an example, and comparison, LSTM labels the AQI on May 13, 2023, at 02:00, while XGBoost does not. Based on the value details, the AQI changed by 33 points from 86 (moderate) at 01:00 to 119 (unhealthy for sensitive group) at 02:00. Then, on May 10, 2019, at 07:00, XGBoost captured the AQI value as an anomaly, while LSTM did not. At that time, the AQI was 8 (good), and an hour earlier, it was 23 (good). Based on these conditions, LSTM is better at capturing extreme changes in data values, while XGBoost is better at capturing extreme data values.

Figure 6 shows the results of AQI anomalies detected by LSTM as the best method based on validation and classification evaluation, which occurred between 21:00 to 09:00. When related to the explanatory variables used, the wind speed (WS10M), rainfall (PRECTOTCORR), and air humidity (RH2M) variables show considerable variability compared to other variables. Based on the exploration, there is a noticeable difference in values between daytime and nighttime. At night, the average wind speed is under 3 m/s, ranging from 2.49 m/s (23:00) to 2.95 m/s (07:00). In comparison, at other times, it ranges from 3.07 m/s (17:00) to 3.64 m/s (15:00). Furthermore, the corrected rainfall at night ranged from 0.25 mm (22:00) to 0.34 mm (03:00). Also, it affected the air humidity, which averaged above 85% from 19:00 to 06:00, while at other times tended to be under this value. During the hours influenced by the three weather variables, around 20:00 to 04:00, the AQI detected as an anomaly tends to increase (for example, a change in the moderate category to unhealthy for sensitive groups). This result is consistent with the data exploration. In line with Kusumaningtyas *et al*. [31], who mentioned that at night, the wind speed tends to be low and calm (<3 m/s) so that surface pollutants cannot move vertically in the inversion layer that surrounds the surface air at night like a "serving hood". In addition, a high AQI is in line with the high amount of human activity during peak hours, mainly due to traffic and vehicle pollution [32].

Then, when viewed on a monthly timeframe, anomalies tend to occur in November-March. Exploration of explanatory variables shows that wind speed, rainfall, and humidity are three variables with high variability, such as hourly timeframes. However, over a more extended period, climate variables such as rainfall will be more visible than weather conditions such as wind speed due to the influence of seasonality. During the rainy season, around November-March, the rainfall correction ranges from 0.29 mm (November and April) to 0.65 mm (February). This value is higher than during the dry season (May-October), which ranges from 0.06 mm (August) to 0.19 mm (May). High rainfall will cause an increase in air humidity, such as in February, with an average humidity of 85.63%. This condition follows the research of Cholianawati *et al*. [33], with data from 2021 showing that there is a monsoonal rainfall pattern with one peak in December-January-February. Then, according to the Indonesia Meteorological, Climatological, and Geophysical Agency (BMKG) in Yulinawati *et al*. [34], Jakarta entered the rainy season in November, causing wet precipitation, which decreased the AQI value. In addition, according to Istiana *et al*. [32], the La Nina phenomenon results in increased rainfall and lower temperatures, as well as wind speed during the dry season in July-August, which can affect the AQI. Low temperature and wind cause pollutants to be trapped and form particles, thus increasing the AQI. This condition is in line with the average of the data used, which is 27.02 °C (July) and 27.05 °C (August), lower than most months in the rainy season, such as in November (28.11 °C).

(a)

(b)

Figure 5. Number of anomalies by (a) LSTM and (b) XGBoost method

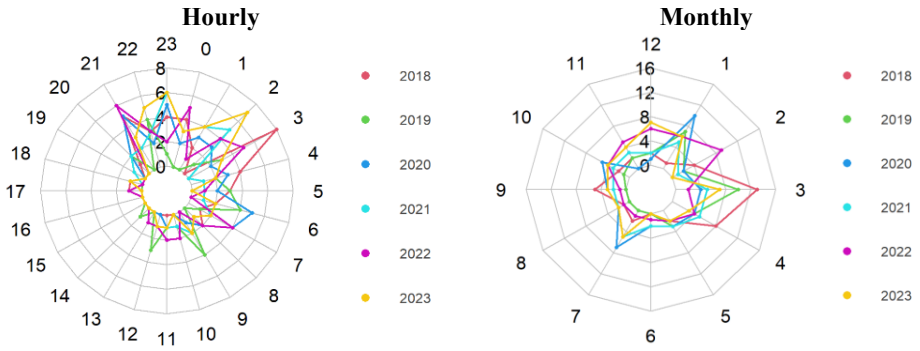

Figure 6. Number of anomalies detected by the best method (LSTM) based on hour and month

4. CONCLUSION

The results of validation and classification of the AQI show similar results, i.e., the LSTM method is better than the XGBoost method. LSTM performs best with the combination of MR (2), 4-sigma, and WFE to perform anomaly detection especially in extreme value changes, while in XGBoost, the best combination is MR (3), 4-sigma, and FE to capture extreme values. However, these results need to be studied further as various factors affect the model evaluation, especially labeling the actual data. The anomaly detection results in the AQI show that many anomalies occur between 21:00 and 09:00 and in the rainy season. AQI anomalies are inseparable from human activities, industrial activities, and weather conditions (rainfall, humidity, and wind speed). All three influence each other, so improving air quality requires the cooperation of various parties and the formulation of appropriate policies. This study has limitations related to the labeling process and modeling approach used. Future research could explore alternative labeling schemes, such as the use of different MR range thresholds (e.g., MR >3), which may provide higher sensitivity in detecting changes in diversity. Other research directions that could be developed from the labeling side include ranking-based labeling, distance-based measures, or probability-based labeling. Additionally, predictive performance could be improved by exploring hybrid modeling strategies, where LSTM and XGBoost are combined with each other or with advanced architectures such as variational autoencoders (VAE–LSTM) to capture short-term information and complex anomaly characteristics, especially in AQI data.

FUNDING INFORMATION

This research was supported by grants from the BIMA Program of the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia (currently the Ministry of Higher Education, Science, and Technology) under contract number 027/E5/PG.02.00.PL/2024.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Muhammad Rizky Nurhambali		✓	✓		✓	✓	✓		✓	✓			✓	
Yenni Angraini	✓	✓		✓			✓			✓	✓	✓		✓
Anwar Fitrianto	✓	✓		✓				✓		✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of the research are available in AirNow and NASA POWER site at https://www.airnow.gov/ and https://power.larc.nasa.gov/, reference [21], [22]. These data were derived from the following resources available in the public domain: https://power.larc.nasa.gov/data-access-viewer/. But unfortunately, the specific URL for reference [21] is no longer available. To support data accessibility and reproducibility, the dataset used in this study has also been made publicly available through at https://drive.google.com/drive/folders/17-eqZRUwLQvkk7iHPCZdhsQoLnNVNQVQ.

REFERENCES

[1] H. Borges, R. Akbarinia, and F. Masseglia, "Anomaly detection in time series," in *Transactions on Large-Scale Data- and Knowledge-Centered Systems L*, Springer Berlin Heidelberg, 2021, pp. 46–62, doi: 10.1007/978-3-662-64553-6_3.

[2] K. Choi, J. Yi, C. Park, and S. Yoon, "Deep learning for anomaly detection in time-series data: review, analysis, and guidelines," *IEEE Access*, vol. 9, pp. 120043–120065, 2021, doi: 10.1109/ACCESS.2021.3107975.

[3] Z. Fang, S. Yang, C. Lv, S. An, and W. Wu, "Application of a data-driven XGBoost model for the prediction of COVID-19 in the USA: a time-series study," *BMJ Open*, vol. 12, no. 7, Jul. 2022, doi: 10.1136/bmjopen-2021-056685.

[4] S. Lin and Y. Feng, "Research on stock price prediction based on orthogonal Gaussian basis function expansion and Pearson correlation coefficient weighted LSTM neural network," *Advances in Computer, Signals and Systems*, vol. 6, no. 5, 2022, doi: 10.23977/acss.2022.060504.

[5] R. Ghosh, K. Causey, K. Burkart, S. Wozniak, A. Cohen, and M. Brauer, "Ambient and household PM2.5 pollution and adverse perinatal outcomes: a meta-regression and analysis of attributable global burden for 204 countries and territories," *PLOS Medicine*, vol. 18, no. 9, Sep. 2021, doi: 10.1371/journal.pmed.1003718.

[6] B. Haryanto, B. Jalaludin, A. Asyary, N. Roestandy, and F. Nugraha, "Associations between ambient PM2.5 levels and children's pneumonia and asthma during the COVID-19 pandemic in Greater Jakarta (Jabodetabek)," *Annals of Global Health*, vol. 91, no. 1, Feb. 2025, doi: 10.5334/aogh.4623.

[7] T. Rosalie, Y. Riani, and Meiryani, "The urgency of implementing carbon tax on Jakarta's air quality (tax incentives as moderation variable)," *E3S Web of Conferences*, vol. 559, Aug. 2024, doi: 10.1051/e3sconf/202455904025.

[8] S. Alla and S. K. Adari, "Practical use cases of anomaly detection," in *Beginning Anomaly Detection Using Python-Based Deep Learning*, Berkeley, United States: Apress, 2019, pp. 297–318, doi: 10.1007/978-1-4842-5177-5_8.

[9] X. Wang, T. Zhao, H. Liu, and R. He, "Power consumption predicting and anomaly detection based on long short-term memory neural network," in *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, Apr. 2019, pp. 487–491, doi: 10.1109/ICCCBDA.2019.8725704.

[10] Y. Liu, "Anomaly detection in multivariate time series using ensemble method." M.Eng. thesis, School of Computer Science and Engineering, Nanyang Technological University, Singapore, 2021, doi: 10.32657/10356/155731.

[11] P. Trizoglou, X. Liu, and Z. Lin, "Fault detection by an ensemble framework of extreme gradient boosting (XGBoost) in the operation of offshore wind turbines," *Renewable Energy*, vol. 179, pp. 945–962, Dec. 2021, doi: 10.1016/j.renene.2021.07.085.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[13] J. Zhang, Y. Zeng, and B. Starly, "Recurrent neural networks with long term temporal dependencies in machine tool wear diagnosis and prognosis," *SN Applied Sciences*, vol. 3, no. 4, Apr. 2021, doi: 10.1007/s42452-021-04427-5.

[14] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022, doi: 10.1109/ACCESS.2022.3152828.

[15] X. Wang, Y. Yang, X. Zhao, M. Huang, and Q. Zhu, "Integrating field images and microclimate data to realize multi-day ahead forecasting of maize crop coverage using CNN-LSTM," *International Journal of Agricultural and Biological Engineering*, vol. 16, no. 2, pp. 199–206, 2023, doi: 10.25165/j.ijabe.20231602.7020.

[16] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[17] Z. Arif Ali, Z. H. Abduljabbar, H. A. Tahir, A. B. Sallow, and S. M. Almufti, "Extreme gradient boosting algorithm with machine learning: a review," *Academic Journal of Nawroz University*, vol. 12, no. 2, pp. 320–334, May 2023, doi: 10.25007/ajnu.v12n2a1612.

[18] C. Bentéjac, A. Csörgő, and G. M.-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

[19] S. Bej, N. Davtyan, M. Wolfien, M. Nassar, and O. Wolkenhauer, "LoRAS: an oversampling approach for imbalanced datasets," *Machine Learning*, vol. 110, no. 2, pp. 279–301, Feb. 2021, doi: 10.1007/s10994-020-05913-4.

[20] S. Pramana, D. Y. Paramartha, Y. Adhinugroho, and M. Nurmalasari, "Air pollution changes of Jakarta, Banten, and West Java, Indonesia during the first month of COVID-19 pandemic," *Journal of Business, Economics and Environmental Studies*, vol. 10, no. 4, pp. 15–19, 2020, doi: 10.13106/jbees.2020.vol10.no4.15.

[21] United States Environmental Protection Agency, "Maps and data U.S. embassies and consulates," *airnow.gov*. Accessed: Jun. 16, 2024. [Online]. Available: https://drive.google.com/drive/folders/17-eqZRUwLQvkk7iHPCZdhsQoLnNVNQVQ

[22] NASA Prediction of Worldwide Energy Resources (POWER), "Data access viewer (DAV)," *power.larc.nasa.gov*. Accessed: Jun. 16, 2024. [Online]. Available: https://power.larc.nasa.gov/data-access-viewer

[23] A. Attaallah and R. A. Khan, "SMOTEDNN: a novel model for air pollution forecasting and AQI classification," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 1403–1425, 2022, doi: 10.32604/cmc.2022.021968.

[24] S. Zheng, Y. Fu, Y. Sun, C. Zhang, Y. Wang, and E. Lichtfouse, "High resolution mapping of nighttime light and air pollutants during the COVID-19 lockdown in Wuhan," *Environmental Chemistry Letters*, vol. 19, no. 4, Mar. 2021, doi: 10.1007/s10311-021-01222-x.

[25] Y. Tian, X. A. Yao, L. Mu, Q. Fan, and Y. Liu, "Integrating meteorological factors for better understanding of the urban form-air quality relationship," *Landscape Ecology*, vol. 35, no. 10, pp. 2357–2373, Oct. 2020, doi: 10.1007/s10980-020-01094-6.

[26] T. Handhayani, "An integrated analysis of air pollution and meteorological conditions in Jakarta," *Scientific Reports*, vol. 13, no. 1, Apr. 2023, doi: 10.1038/s41598-023-32817-9.

[27] M. S. Rahman, A. H. Chowdhury, and M. Amrin, "Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh," *PLOS Global Public Health*, vol. 2, no. 5, May 2022, doi: 10.1371/journal.pgph.0000495.

[28] C.-X. Lv, S.-Y. An, B.-J. Qiao, and W. Wu, "Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model," *BMC Infectious Diseases*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12879-021-06503-y.

[29] W. H. Woodall and D. C. Montgomery, "Using ranges to estimate variability," *Quality Engineering*, vol. 13, no. 2, pp. 211–217, Dec. 2000, doi: 10.1080/08982110108918643.

[30] J. Zheng, D. Feng, Z. Yang, Y. Xiang, H. Zhang, and S. Li, "TransKS: an anomaly detection method for telecommunication networks based on deep learning," *IEEE Access*, vol. 11, pp. 118048–118060, 2023, doi: 10.1109/ACCESS.2023.3326815.

[31] S. D. A. Kusumaningtyas, A. N. Khoir, E. Fibriantika, and E. Heriyanto, "Effect of meteorological parameter to variability of particulate matter (PM) concentration in urban Jakarta city, Indonesia," *IOP Conference Series: Earth and Environmental Science*, vol. 724, no. 1, Apr. 2021, doi: 10.1088/1755-1315/724/1/012050.

[32] T. Istiana *et al.*, "Causality analysis of air quality and meteorological parameters for PM2.5 characteristics determination: evidence from Jakarta," *Aerosol and Air Quality Research*, vol. 23, no. 9, 2023, doi: 10.4209/aaqr.230014.

[33] N. Cholianawati *et al.*, "Diurnal and daily variations of PM2.5 and its multiple-wavelet coherence with meteorological variables in Indonesia," *Aerosol and Air Quality Research*, vol. 24, no. 3, 2024, doi: 10.4209/aaqr.230158.

[34] H. Yulinawati, T. Khairani, and L. Siami, "Analysis of indoor and outdoor particulate (PM 2.5) at a women and children's hospital in West Jakarta," *IOP Conference Series: Earth and Environmental Science*, vol. 737, no. 1, Apr. 2021, doi: 10.1088/1755-1315/737/1/012067.

BIOGRAPHIES OF AUTHORS

Muhammad Rizky Nurhambali is a student pursuing a master's degree in Statistics and Data Science at IPB University. He received his Bachelor of Statistics degree in 2019 from the Department of Statistics, IPB University, Indonesia. During his education, he was active as a teaching assistant in several courses at the undergraduate level, such as statistical methods and time series forecasting methods, and at the master level in statistics for agricultural and biological sciences. His research areas of interest include regression, time series, and machine learning. He can be contacted at email: rizkynurhambali@apps.ipb.ac.id or rizky2710.edu@gmail.com.

Yenni Angraini is an academic journey led her to excel in statistics, earning her Bachelor's, Master's, and Doctorate degrees in Statistics from IPB University, currently serving as a dedicated faculty member. She is a respected lecturer at the Study Program of Statistics and Data Science, School of Data Science, Mathematics, and Informatics, IPB University. Her commitment to education extends beyond the classroom, as evidenced by her active involvement in professional organizations. She is a proud member of the Ikatan Statistisi Indonesia (Indonesian Statistician Association) and the Forum Pendidikan Tinggi Statistika Indonesia (Indonesian Higher Education Statistics Forum). She can be contacted at email: y_angraini@apps.ipb.ac.id.

Anwar Fitrianto is currently active as a lecturer at the Study Program of Statistics and Data Science, School of Data Science, Mathematics, and Informatics, IPB University. He obtained his bachelor's degree in Statistics in 1999 from the Department of Statistics, IPB University, Indonesia. Both master of science and Ph.D. in statistics were obtained in 2005 and 2010, respectively, from Universiti Putra Malaysia. He has strong expertise in robust statistics, statistics modeling, experimental design, and statistical process control. He can be contacted at email: anwarstat@gmail.com.