Institut Riset dan Publikasi Indonesia (IRPI)
IJATIS: Indonesian Journal of Applied Technology and Innovation Science
Journal Homepage: https://journal.irpi.or.id/index.php/ijatis
Vol. 2 Iss. 2 August 2025, pp: 63-69
ISSN(P): 3032-7466 | ISSN(E): 3032-7474

PM 2.5 Prediction Using the Long Short-Term Memory Algorithm

Syaid El Hasyim[1*], Nurazizah[2], Muhammad Yudha Pratama[3], Umairah Rizkya Gurning[4], Batrisia Khairunnisa[5]

[1,2,3,4]Department of Information Systems, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[5]Department of Architecture, Faculty of Architecture, Haliç University, Turkiye

E-Mail: [1]12050312518@students.uin-suska.ac.id, [2]12050321684@students.uin-suska.ac.id, [3]12050312952@students.uin-suska.ac.id, [4]11950320687@students.uin-suska.ac.id, [5]batrisiakh@gmail.com

Received Nov 28th 2024; Revised Jul 26th 2025; Accepted Aug 20th 2025; Available Online Aug 30th 2025
Corresponding Author: Syaid El Hasyim

Abstract

Air pollution poses a serious threat to human health and the environment, with far-reaching impacts on various aspects of life. Among its most harmful components is particulate matter less than 2.5 micrometers in diameter (PM2.5), which contributes significantly to degraded air quality. Accurate prediction of PM2.5 concentrations is crucial for protecting public health and informing policy-making. This study employs the Long Short-Term Memory (LSTM) algorithm, a deep learning method well-suited for modeling large, complex, and time-dependent datasets, to forecast PM2.5 levels in Delhi, India. The dataset comprises daily records from January 1, 2015, to July 1, 2020. The proposed model achieved a Mean Absolute Percentage Error (MAPE) of 25.22%, indicating moderate predictive accuracy. These results demonstrate that the LSTM algorithm can serve as an effective tool for forecasting PM2.5 concentrations, providing valuable insights for air quality management and environmental planning.

Keywords: Air Pollution, Long Short-Term Memory, Mean Absolute Percentage Error, PM 2.5

1. INTRODUCTION

In recent decades, air pollution has emerged as a critical global issue, affecting both environmental quality and human health. One of the most hazardous pollutants is fine particulate matter (PM2.5), which consists of microscopic particles with a diameter of less than 2.5 micrometers. According to M. Ja'far Sodiq [1], air pollution refers to the release of chemicals, energy, or other substances into the atmosphere as a result of human activities, which reduces air quality and poses potential health hazards. Factors such as population growth, congested transportation systems, rapid infrastructure expansion, and industrial activities have been identified as major contributors to deteriorating air quality worldwide [2]. Given its ability to penetrate deep into the respiratory system and its proven association with severe health risks, PM2.5 has become a primary focus in environmental monitoring and public health studies.

Several studies have noted that air quality measurements are inherently time series data, allowing predictions to be made based on historical trends [3]. The time series method utilizes data collected sequentially over time to identify patterns and forecast future values [4]. High concentrations of PM2.5 particles are a key indicator of air pollution and can have significant social, economic, and public health consequences [5]. The adverse effects of fine particulate matter have been recognized as a leading cause of public health deterioration [6]. Therefore, accurate PM2.5 forecasting is essential for public welfare, enabling timely warnings for outdoor activities, informing adjustments to public transportation policies, and supporting environmental conservation efforts [7].

Advancements in artificial intelligence, particularly deep learning, have enabled the modeling of nonlinear dependencies, seasonal variations, and sequential patterns in air quality data [8]. Among the available techniques, the Long Short-Term Memory (LSTM) network has proven effective for processing time series data with long intervals and delayed dependencies [9]. Previous studies have reported strong predictive performance of LSTM models, with metrics such as Nash-Sutcliffe Efficiency (NSE) ranging from 0.86 to 0.94 and correlation coefficients (CC) from 0.93 to 0.97 in air quality predictions [10]. Furthermore, applications of LSTM and related models, such as LSTM-RNN and GRU, have achieved high accuracy in

other complex domains, including eukaryotic DNA exon prediction, with results reaching 96.1% accuracy and a reduction in training time [11]. Building on this evidence, the present study applies the LSTM algorithm to forecast PM2.5 concentrations using historical air quality data including pollutant levels, humidity, temperature, and other relevant factors with the goal of developing a reliable predictive model to support air quality management.

The LSTM algorithm is used in this study to predict air quality using PM2.5. It is hoped that a model will be created that can effectively predict air quality using historical data on air characteristics, such as pollution concentration, humidity, temperature, and other relevant factors.

2. MATERIAL AND METHOD

This study uses a LSTM algorithm intended to provide predictions regarding air quality using PM2.5 in India based on historical data. The stages of the research can be seen in Figure 1.

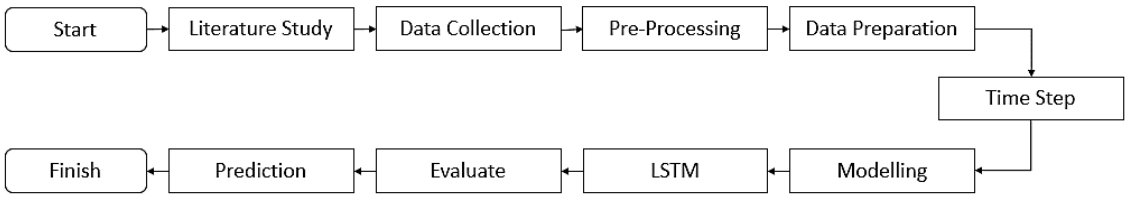

Figure 1. Research Methodology

The following steps will be carried out in this study to forecast the PM2.5 pollutant model:

2.1. Study Literature

2.1.1. PM 2.5

The tiny airborne particle known as PM 2.5 has been linked to a number of diseases [12]. Due to their small size (2.5 m in diameter) and their ability to absorb poisons, these particles can infiltrate the inner lungs and cause various diseases [13]. Other pollutants can chemically react with each other in the atmosphere to produce PM2.5. Motor vehicle fumes, industrial fumes, and wood burning fumes are some of the pollutants that can produce PM2.5 [14].

PM2.5, PM10, NO, and other contaminants are used to measure air pollution. PM 2.5 is the term used to describe small diameter fine particles found in the environment. It is also affected by meteorological factors such as wind, precipitation, temperature, etc. Due to changes in concentration, they have no direct effect on the Air Quality Index (AQI). Environmental authorities use the AQI, a significant measure, to report air quality. Predictive analysis is based on the severity displayed compared to others [15].

2.1.2. Deep Learning

Deep learning is increasingly important as a technique for data analysis along with the development of big data technology and artificial intelligence [16]. Deep learning has the ability to extract powerful features from data and is data-driven. Convolutional Neural Networks (CNN), Gated Recurrent Unit (GRU), Recursive Neural Network (RNN), and LSTM models are some examples of deep learning-based models that are increasingly being used in time series forecasting. According to Bi et al. (2023), these models are often used for predictions in various disciplines, including path prediction, wind speed prediction, water level prediction, signal prediction, etc [17].

In the study example [18], the level of air pollution is predicted using the LSTM. Repetitive artificial neural networks using LSTM can solve long-term learning problems. LSTM is used in research[18] to identify patterns and relationships between weather and air pollution levels during the previous hour. Air pollution levels for the next hour can be predicted using weather and air pollution data from the previous hour. Air quality prediction in Visakhapatnam is the focus of the case study [19] with LSTM. LSTM is used to model and estimate air pollution concentrations periodically in the investigation [19]. In estimating the level of air pollutant concentration in Visakhapatnam, LSTM was used to yield remarkable results.

2.1.3. Long Short-Term Memory (LSTM)

Time series models can be used to explain the dynamic and continuous nature of PM2.5 concentrations over time. In recent years, LSTM has become one of the most widely used time series prediction models [20]. A unique variation of the RNN is the LSTM neural network. The capacity to incorporate both long-term and short-term dependencies into the input sequence is its main characteristic, hence the name [21]. To handle events with long intervals and significant delays in the time series,

Hochreiter and Schmidhuber (1997) developed a Long-Short-Term Memory (LSTM) neural network, which was developed from the RNN artificial neural network [22].

The basic principle of an LSTM network is to store a cell to store state information on each neuron in the network and to set up three logic gates - the input gate (it), the forget gate (ft), and the output gate (ot) - to control adding or deleting data from cell memory. According to Karyadi and Santoso [9], [23], these gates allow the LSTM to function reliably over a long period of time without experiencing gradient burst or gradient loss. Figure 2 illustrates the internal structure of the LSTM [24].

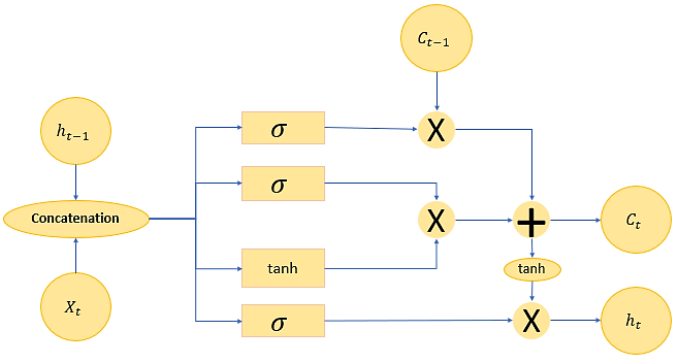

Figure 2. LSTM Internal Structure

Ct represents the memory of the current time, xt represents the input of the current time, and ht represents the cell state value at the current time. The LSTM calculation formula is as follows [25]:

1. Forget gate calculation formula:

$$f_t = \sigma(IN_f . [h_{t-1}, x_t] + b_f) \quad (1)$$

2. Gate input calculation formula:

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

$$\tilde{C}_t = \tanh(IN_c . [h_{t-1}, x_t] + b_c) \quad (3)$$

3. The gate output calculation formula:

$$O_t = \sigma(IN_O . [h_{t-1}, x_t] + b_O) \quad (4)$$

$$h_t = O_t * \tanh(C_t) \quad (5)$$

4. Updating unit status:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t) \quad (6)$$

Bi, bf, bc, and bo are input, forget, update, and output gates, respectively, while Wi, Wf, WC, and Wo are weight matrices for the corresponding input, forget, update, and output gates. Calculations are performed to determine the ht output at the current t time and the updated state of the cell Ct at the current t time.

2.2. Data Collection

Data in CSV (Comma Separated Value) format in this study were obtained from Kaggle at "https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-India". The data consists of 26 cities in India and 14 parameters. In this study, the data used were PM2.5 from the city of Delhi, which covered the time period from 1 January 2015 to 1 July 2020. Table 1 displays the resulting dataset.

2.3. Pre-Processing

The normalization stage of the dataset begins with changing the attribute data type date, whose type is object, to datetime. After that, the process is carried outdata cleaning by checking whether there are still empty rows or not. It turns out that the dataset has 2 empty rows which are then filled with commandsfeel (blank data filled with average value).

Table 1. India Country Dataset

City	Date	PM2.5	PM10	NO
Ahmedbab	2015-01-01	NaN	NaN	0,92
Ahmedbab	2015-01-02	NaN	NaN	0,97
Ahmedbab	2015-01-03	NaN	NaN	17,4
Ahmedbab	2015-01-04	NaN	NaN	1,7
Ahmedbab	2015-01-05	NaN	NaN	22,1
.....	.....	.....	.....	.....
Visakhapatnam	2020-06-27	15,02	50,94	0,33
Visakhapatnam	2020-06-28	0,38	74,09	3,43
Visakhapatnam	2020-06-29	0,97	65,73	4,45
Visakhapatnam	2020-06-30	0,71	49,97	4,05
Visakhapatnam	2020-07-01	15	66	0,04

2.4. Preparation Data

The research dataset is divided using the hold-out method into three, namely the period January 1, 2015, to June 31, 2019, for training data, July 1, 2019, to December 31, 2019, for validation data, and January 1, 2020, to July 1, 2020, for data testing.

2.5. Training and Testing Process

This LSTM model uses two hidden layers with a batch size of 32, an epoch of 800, and RMSprop as an optimizer before training. Mean Absolute Percentage Error (MAPE) is the unit of measurement used. To determine how effective the model used is in the training process, the validation step is carried out after the training phase. The learning model obtained during the previous training procedure will be regenerated in this step. The testing part of this method is used to evaluate how well the LSTM model works.

3. RESULTS AND DISCUSSION

Dataset taken from kaggle "https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india", which is air quality data in India in the period 2015 – 2020. From this data there are 14 Parameters and 23 Cities in India. From these data, only 1 parameter was modeled and predicted, namely PM 2.5 in the city of Delhi, India. The first step is to convert to Datetime. The procedure for calculating characteristic quartiles is the next step, after which any outliers are interpolated. Table 2 displays the original dataset, while Table 3 displays the data to be used.

3.1. PM2.5 prediction

To predict PM2.5, PM2.5 parameters are used in Delhi, India. The following is a comparison of the plotting results against the original dataset in the image, with the data that has been transformed can be seen in Figure 3. PM 2.5 prediction process with LSTM, data sharing is carried out using the Hold Out technique, where training data starts from 1 January 2015 to 31 June 2019, validation data from 1 July 2019 to 31 December 2019 and data testing from 1 January 2020 to 1 July 2020.

The results of the data distribution are used for predictions with the LSTM algorithm. From the results processing data, the results are obtained with a graph that can be seen in Figure 4.

Table 2. India Country Dataset

City	Date	PM2.5	PM10	NO
Ahmedbab	2015-01-01	NaN	NaN	0,92
Ahmedbab	2015-01-02	NaN	NaN	0,97
Ahmedbab	2015-01-03	NaN	NaN	17,4
Ahmedbab	2015-01-04	NaN	NaN	1,7
Ahmedbab	2015-01-05	NaN	NaN	22,1
.....	.....	.....	.....	.....
Visakhapatnam	2020-06-27	15,02	50,94	0,33
Visakhapatnam	2020-06-28	0,38	74,09	3,43
Visakhapatnam	2020-06-29	0,97	65,73	4,45
Visakhapatnam	2020-06-30	0,71	49,97	4,05
Visakhapatnam	2020-07-01	15	66	0,04

Table 3. Dataset Used

City	Date	PM2.5	PM10	NO
Delhi	2015-01-01	313,22	607,98	69,16
Delhi	2015-01-02	186,18	269,55	62,09
Delhi	2015-01-03	87,18	131,9	25,73

City	Date	PM2.5	PM10	NO
Delhi	2015-01-04	151,84	241,84	25,01
Delhi	2015-01-05	146,6	219,13	14,01
…….	…….	…….	…….	…….
Delhi	2020-06-27	39,8	155,94	10,88
Delhi	2020-06-28	59,52	308,65	12,67
Delhi	2020-06-29	44,86	184,12	10,5
Delhi	2020-06-30	39,8	91,98	5,99
Delhi	2020-07-01	54,01	128,66	6,33

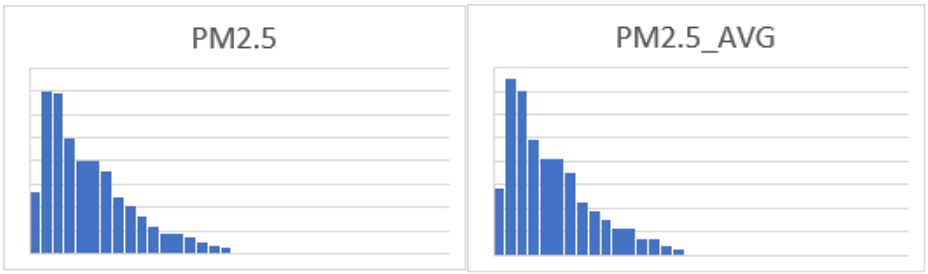

Figure 3. Comparison of Original and Transformed PM 2.5 Data

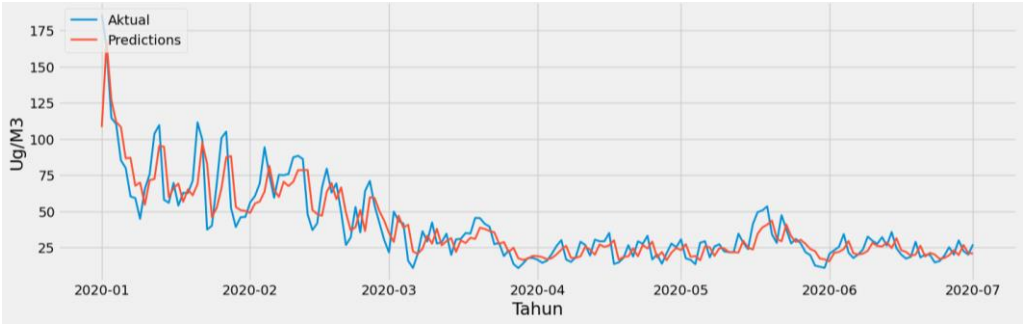

Figure 4. Graph of PM2.5 Prediction Results

Figure 4 shows the prediction results of PM2.5 in the city of Delhi, India, which produces good prediction accuracy, where you can see the prediction results on the red graph which are almost the same as the original blue data. The prediction results using LSTM can also be seen in Table 4.

Table 4. PM2.5 Prediction Results

Date	PM2.5_AVG	Predictions
2020-01-01	186,07	108,1359558
2020-01-02	163,52	166,9525452
2020-01-03	114,45	126,8785248
2020-01-04	110,3	111,7203674
2020-01-05	85,105	108,1705551
…….	…….	…….
2020-06-27	19,9	22,94326591
2020-06-28	29,76	19,61851311
2020-06-29	22,43	26,55563164
2020-06-30	19,9	20,96610069
2020-07-01	27,005	20,7005806

Table 4 presents the PM 2.5 prediction results for the City of Delhi, India, using the LSTM, yielding nearly identical results. This matter is demonstrated in Table 5, which shows an accuracy value of 25.22% using the MAPE metric. The accuracy value for predictions using LSTM indicates that the obtained prediction results are quite good.

3.2. Discusions

Despite numerous studies on air quality prediction, many focus on short-term forecasting or employ statistical models with limited capacity to capture nonlinear patterns. Prior works often overlook the integration of multiple environmental variables beyond pollutant concentrations. This study addresses that gap by applying an LSTM model to a multi-year dataset from Delhi, India, incorporating meteorological

factors such as humidity and temperature. By doing so, it advances the predictive capability for PM2.5 concentrations, offering more robust insights than models relying solely on historical pollutant data.

The primary strength of this research lies in its use of the LSTM algorithm, which effectively models long-term dependencies in time series data. The dataset spans over five years, allowing the model to learn from extensive historical patterns. Achieving an MAPE of 25.22% demonstrates moderate predictive accuracy, validating the model's suitability for PM2.5 forecasting. Furthermore, the integration of meteorological variables enhances prediction reliability, while the transparent reporting of model parameters facilitates reproducibility and comparability with future research.

Future studies could enhance model performance by exploring hybrid approaches that combine LSTM with other machine learning or statistical techniques. Expanding the dataset to include additional socio-economic and industrial activity indicators may further improve prediction accuracy. Testing the model in different geographical regions would assess its generalizability, while integrating real-time data streams could enable dynamic forecasting systems. Additionally, optimizing hyperparameters through automated search techniques, such as Bayesian optimization, could reduce error rates and provide more reliable early warnings for public health and environmental management.

4. CONCLUSION

This study developed and applied a LSTM model for forecasting PM2.5 concentrations in Delhi, India, using daily air quality data from January 1, 2015, to July 1, 2020. The model, configured with two hidden layers, a batch size of 32, the RMSprop optimizer, and 800 training epochs, achieved a MAPE of 25.22%, indicating moderate predictive accuracy. These findings demonstrate the potential of LSTM as an effective tool for modeling complex time series data in environmental applications. The model's predictive capability can assist in issuing timely public health advisories, guiding transportation policies, and supporting environmental management strategies. Future research could explore the integration of additional meteorological and socio-economic variables, the application of hybrid models, and testing across different geographic regions to further enhance predictive performance.

REFERENCES

[1] E. I. S. M. Ja'far Sodiq, "Comparison of Naive Bayes and K-Nearest Neighbor Methods for Air Classification in DKI Jakarta," Yogyakarta, Dec. 2019.

[2] K. Auliasari, M. Kertaningtyas, and J. Raya Karanglo Km, "Air Quality Analysis Using the K-Means Algorithm," 2021. [Online]. Available: http://e-journal.stmiklombok.ac.id/index.php/jireISSN.2620-6900

[3] A. Khumaidi, R. Raafi, I. Permana Solihin, and J. Rs Fatmawati, "Testing the Long Short Term Memory Algorithm for Predicting Air Quality and Temperature in the City of Bandung,"Telematics Journal, flight. 15, no. 1, 2020.

[4] Lingga Yuliana, "Analysis of Sales Planning Using the Time Series Method (Case Study at PD. Sumber Jaya Aluminum),"Management Partner (JMM Online), vol. 3, pp. 780–789, Jul. 2019.

[5] Z. Zhang, Y. Zeng, and K. Yan, "A hybrid deep learning technology for PM2.5 air quality forecasting," Environmental Science and Pollution Research, vol. 28, no. 29, pp. 39409–39422, Aug. 2021, doi: 10.1007/s11356-021-12657-8.

[6] T. Xueet al., "Rapid improvement of PM2.5 pollution and associated health benefits in China during 2013–2017," Sci China Earth Sci, vol. 62, no. 12, pp. 1847–1856, Dec. 2019, doi: 10.1007/s11430-018-9348-2.

[7] N. S. Muruganandam and U. Arumugam, "Seminal Stacked Long Short-Term Memory (SS-LSTM) Model for Forecasting Particulate Matter (PM2.5 and PM10)," Atmosphere (Basel), vol. 13, no. 10, Oct. 2022, doi: 10.3390/atmos13101726.

[8] G. I. Drewil and R. J. Al-Bahadili, "Air pollution prediction using LSTM deep learning and metaheuristics algorithms," Measurement: Sensors, vol. 24, Dec. 2022, doi: 10.1016/j.measen.2022.100546.

[9] B. Liu, Z. Yu, Q. Wang, P. Du, and X. Zhang, "Prediction of SSE Shanghai Enterprises index based on bidirectional LSTM model of air pollutants," Expert Syst Appl, vol. 204, Oct. 2022, doi: 10.1016/new.2022.117600.

[10] M. Krishan, S. Jha, J. Das, A. Singh, M. K. Goyal, and C. Sekar, "Air quality modelling using long short-term memory (LSTM) over NCT-Delhi, India," Air Qual Atmos Health, vol. 12, no. 8, pp. 899–908, Aug. 2019, doi: 10.1007/s11869-019-00696-7.

[11] P. J. Canatalay and O. N. Ucan, "A Bidirectional LSTM-RNN and GRU Method to Exon Prediction Using Splice-Site Mapping," Applied Sciences (Switzerland), vol. 12, no. 9, May 2022, doi: 10.3390/app12094390.

[12] Q.: Journalet al., "Dust Respirable Concentration 'Particulate Matter' (Pm2.5) And Health Disorders Communities In Settlement Around Electric Steam Power Plant," PROMOTIVE, 2019.

[13] T. Meidya and R. Yudhastuti, “Literature Review: Long-term Exposure to PM2.5 is at Risk of Increasing Mortality Due to COVID-19,” 2021.
[14] M. Unik and Sri Nadriati, “Overview: Random Forest Algorithm for PM2.5 Estimation Based on Remote Sensing,” Jurnal CoSciTech (Computer Science and Information Technology), vol. 3, no. 3, pp. 422–430, Dec. 2022, doi: 10.37859/coscitech.v3i3.4380.
[15] D. Saravanan and K. Santhosh Kumar, “Improving air pollution detection accuracy and quality monitoring based on bidirectional RNN and the Internet of Things,” Mater Today Proc, 2022, doi: 10.1016/j.matpr.2021.04.239.
[16] S. Roy et al., “Deep Learning for Classification and Localization of COVID-19 Markers in Point-of-Care Lung Ultrasound,” IEEE Trans Med Imaging, vol. 39, no. 8, pp. 2676–2687, Aug. 2020, doi: 10.1109/TMI.2020.2994459.
[17] J. Bi, L. Zhang, H. Yuan, and J. Zhang, “Multi-indicator water quality prediction with attention-assisted bidirectional LSTM and encoder-decoder,” Inf Sci (N Y), vol. 625, pp. 65–80, May 2023, doi: 10.1016/j.ins.2022.12.091.
[18] A. Jain, A. Bhasin, and V. Gupta, “Prediction of air pollution using LSTM-based recurrent neural networks,” 2019. [Online]. Available: http://WorstPolluted.org
[19] K. S. Rao, G. L. Devi, and N. Ramesh, “Air Quality Prediction in Visakhapatnam with LSTM based Recurrent Neural Networks,” International Journal of Intelligent Systems and Applications, vol. 11, no. 2, pp. 18–24, Feb. 2019, doi: 10.5815/ijisa.2019.02.03.
[20] H. Yang, J. He, C. Zhou, and L. Li, “A Particulate Matter 2.5 Concentration Forecasting Method Based on Multi-Input LSTM,” in 2020 IEEE 6th International Conference on Computer and Communications, ICCC 2020, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 1634–1638. doi: 10.1109/ICCC51575.2020.9344921.
[21] S. Ferlito, F. Bosso, S. De Vito, E. Esposito, and G. Di Francia, “LSTM Networks for Particulate Matter Concentration Forecasting,” in Lecture Notes in Electrical Engineering, Springer, 2020, pp. 409–415. doi: 10.1007/978-3-030-37558-4_61.
[22] L. Zhoul, M. Chenl, and Q. Ni, “A hybrid Prophet-LSTM Model for Prediction of Air Quality Index,” in 2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 595–601. doi: 10.1109/SSCI47803.2020.9308543.
[23] Y. Karyadi and H. Santoso, "Air Quality Prediction Using LSTM, Bidirectional LSTM, and GRU Methods,"Journal of Informatics Engineering and Information Systems, flight. 9, no. 1, pp. 671–684, Mar. 2022.
[24] M. P. Ningrum, R. Mutia, H. Azmi, and H. D. Khalifah, “Sentiment Analysis of Twitter Reviews on Google Play Store Using a Combination of Convolutional Neural Network and Long Short-Term Memory Algorithms,” Public Research Journal of Engineering, Data Technology and Computer Science, vol. 2, no. 2, pp. 107–115, Jan. 2025, doi: 10.57152/predatecs.v2i2.1625.
[25] M. F. Fayyad, V. Kurniawan, M. R. Anugrah, B. H. Estanto, and T. Bilal, “Application of Recurrent Neural Network Bi-Long Short-Term Memory, Gated Recurrent Unit and Bi-Gated Recurrent Unit for Forecasting Rupiah Against Dollar (USD) Exchange Rate,” Public Research Journal of Engineering, Data Technology and Computer Science, vol. 2, no. 1, pp. 1–10, Apr. 2024, doi: 10.57152/predatecs.v2i1.1094.