Lafia Journal of Scientific & Industrial Research (LJSIR), Vol. 3(1), 2025
p-ISSN: 3026 – 9288 e-ISSN: 3027 – 1800 pages: 98 – 107
https://lafiascijournals.org.ng/index.php/ljsir/index ⇒ Published by the Faculty of Science,
Federal University of Lafia, Nasarawa State, Nigeria

Air Quality Index Prediction using Deep Learning for Lagos State in Nigeria

Gbenga O. Ogunsanwo[1]*, Phillip T. Odulaja[1], A. A. Omotunde[2] & Olakunle O. Solanke[3]
[1]Department of Computer Science, Tai Solarin University of Education, Ogun State, Nigeria
[2]Department of Computer Science, Babcock University Ilisan Remo, Ogun State, Nigeria
[3]Department of Computer Sciences, Olabisi Onabanjo University, Ago-Iwoye, Nigeria

Abstract

The index for expressing air quality is known as the air quality index (AQI). It could be used to evaluate the effect of air pollution on one's health over a period of time which provides a guide to the community on the adverse health effects of air pollution around them. This paper focused on developing a model for AQI prediction using Deep learning for Lagos state in Nigeria. The study acquired dataset from the OpenWeatherMap API which includes historical air quality and meteorological for Lagos State in Nigeria. Data was preprocessed by handling missing values, converting data types into numerical format using one-hot encoding. The study applied SMOTE technique to ensure balanced dataset. Four distinct Models such as LSTM, CNN, Prophet and SVR were utilized to determine the AQI of Lagos State. The results of balanced datasets used revealed LSTM provides the lowest MSE, RMSE and MAE values of 0.062, 0.249 and 0.149 respectively and higher R^2 value of 0.968 compared with the other model CNN, Prophet and SVR. The paper concluded that in the prediction of the AQI for Lagos State, LSTM outperformed other models such as CNN, Prophet Model and SVR on the validating metrics known as MSE, RMSE, MAE and R^2. The model obtained could be subjected to further research in other geographic regions, such as Delta State or other state by state level analysis which could be expanded to forecast other pollution indices at different levels.

Keywords: AQI, prediction, pollution, deep learning

Article History
Submitted
February 28, 2025
Revised
March 24, 2025
First Published Online
April 2, 2025

Corresponding author
G. O. Ogunsanwo ✉
ogunsanwogo@tasued.edu.ng

doi.org/10.62050/ljsir2025.v3n1.450

Introduction

Advancement in technology has greatly improved our day to day activities but despite its numerous benefits to humanity, it also comes with some challenges especially in the area of the air quality index of an area. The index for reporting air quality is known as the air quality index (AQI). It could be used to evaluate the effect of air pollution on a people's health over some period of time. It has been observed that developed cities are likely to experience more air pollution when comparing with the developing area. Air pollution has become a serious issue in the world [1]. Many urban cities are suffering from air pollution [2]. Air pollution is caused by many pollutants such as Particulate Matter (PM 2.5, PM 10), O3, Acrolein, Asbestos, Benzene, CO, CO2, NO2, NO, SO2, NH3, CS2, Polycyclic Aromatic Hydrocarbons, Synthetic Vitreous Fibers and total Petroleum Hydrocarbons [3]. Due to the challenges caused by air pollution, thus air pollution prediction and monitoring have become hot area that have drawn the attention of many researchers because of its negative impact on the health of the populace as well as on the ecological imbalance it has caused [4].

Nigeria is becoming one of the countries in Africa and the world with the highest level of pollutants in its atmosphere. The country has been facing a lot of challenges with its air quality. There are many reasons why the air quality index is poor. Majorly, it's the result of both global and local factors. The local causes include rapid urbanization, deforestation, industrial activities and poor waste management practices. These have all led to an increase in pollution levels in many parts of Nigeria. The global causes are mainly due to climate change and its effect on rainfall patterns across Nigeria. Climate change has reduced the amount of rainfall which in turn has led to hotter temperatures and drier soil conditions which facilitate the increase in the amount of particles in the air. WHO (World Health Organization) estimated that over seven million people around the globe are affected with one disease or the other as a result of air pollution. The presence of air pollution in any area might increase the chance of people living around that environment contacting disease such as eye diseases, throat infections, lung cancer, asthma, heart issues, skin infections, bronchitis diseases, etc [5]. Sometime, when air pollution is left uncontrolled for a long period it may result in premature mortalities. Children and Pregnant women also suffer great adverse effect of air pollution. The Air quality index (AQI) is usually used to express air quality. The dimensionless factor categorizes air pollution into various quantities. According to European Environment Agency (EEA), the AQI is divided into six different categories from good to hazardous [6].

The mathematical expression to calculate the AQI score is as stated in equation 1 below.

$$P = \frac{P_{high} - P_{low}}{PC_{high} - PC_{low}}(PC - PC_{low}) + P_{low} \quad (1)$$

where: P = AQI; PC = pollutant concentration; PC_{low}= the concentration breakpoint that is ≤ PC

A compressive AQI range for various pollutants is represented in Table 1 for clearer understanding, the level of AQI is in the range 1–3 it's agreed that the presence of pollution is low and quality of air is good in that area. In the second level, the range is fixed into 4–6 it's agreed that the presence of pollution is moderate and the air quality is mentioned as satisfactory. In the third level, the range is fixed into 7–8 it's agreed that the presence of pollution is high and the air quality is for moderate pollution. In the fourth level the poor pollution status in very high with the range of 9.

Table 1: AQI level

Parameter	AQI	SO_2	NO_2	$PM_{2.5}$	PM_{10}	O_3
Low	1	0-88	0-67	0-11	0-16	0-33
Low	2	89-177	68-134	12-23	17-33	34-66
Low	3	178-266	135-200	24-35	34-50	67-100
Moderate	4	267-354	201-267	36-41	52-58	101-120
Moderate	5	355-443	268-334	42-47	59-66	121-140
Moderate	6	444-532	335-400	48-53	67-75	141-187
High	7	533-710	401-467	54-58	84-91	188-213
High	8	711-887	468-534	59-64	84-91	188-213
High	9	888-1064	535-600	65-70	92-100	214-240
Very High	10	>= 1065	>= 601	>= 71	>= 101	>= 241

Source: EEA [6]

Nigeria has the highest burden of mortality from poor air quality in Africa and 4th globally [7]. In Nigeria, rapid urbanization and industrial growth have exacerbated air pollution, particularly in major cities such as Lagos, Kano, and Port Harcourt [8]. The country was ranked 150th out of 180 countries for poor environmental performance index on air quality [9]. Some cities across Nigeria have been prominent to have poor air quality [4, 10–12]. The Air Quality Index (AQI) is seen as a standard metric used globally to report daily air quality levels to the public. It simplifies complex air quality data into a single number and color code that represents the level of health concern. The AQI focuses on health effect that is experienced within a few hours or days after breathing polluted air. This makes it an essential tool for raising public awareness about air quality and its health impacts [6]. Many studies have been carried out in the area of AQI prediction using machine learning such as Srinivasa et al. [13] used support vector regression (SVR), random forest regression (RFR), and CatBoost regression (CR) to determine the AQI of New Delhi, Bangalore, Kolkata, and Hyderabad. The study improved the model performance by comparing the results of imbalanced datasets using synthetic minority oversampling technique (SMOTE) algorithm, the model was validated using RMSE and Accuracy validating metric. The study concluded that random forest regression provides the lowest root mean square error (RMSE) values as well as higher accuracy compared to SVR and CatBoost regression for all the cities. JingyangWang et al [14] carried out the comparison analysis of CNN-ILSTM model on air quality data of Shijiazhuang City, Hebei Province, China from 00:00 on January 1, 2017, to 23:00 on June 30, 2021, employed eight predictive models such as: SVR, RFR, MLP, LSTM, GRU, ILSTM, CNN-LSTM, and CNN-GRU the Model were validated using metrics. The study concluded that CNN-LSTM outperform other models used in the study. Alaba, et al. carried out a study to examine the performance of three data mining classification algorithms: which are Decision Tree, Naïve Bayes and ZeroR to predict the occurrence of air pollution [15]. The study ascertained that Decision tree algorithm yielded the highest prediction accuracy followed by Naive Bayes and ZeroR based on the datasets used. The study recommended that the performance of other classification algorithms could be tested on the SVR. This study developed a comparative analysis of LSTM, CNN, Prophet Model and SVM on Air Quality Index (AQI) of Lagos city in Nigeria.

Materials and Methods

This section explains the approach used for the prediction of AQI. This study employed the use of four algorithms which are: LSTM, CNN, Prophet Model and SVR for the prediction. The procedure includes the following steps: Input Dataset, preprocessing of a dataset, selection, prediction with DL techniques and validation of performance. The overview of this study is represented in Fig. 1.

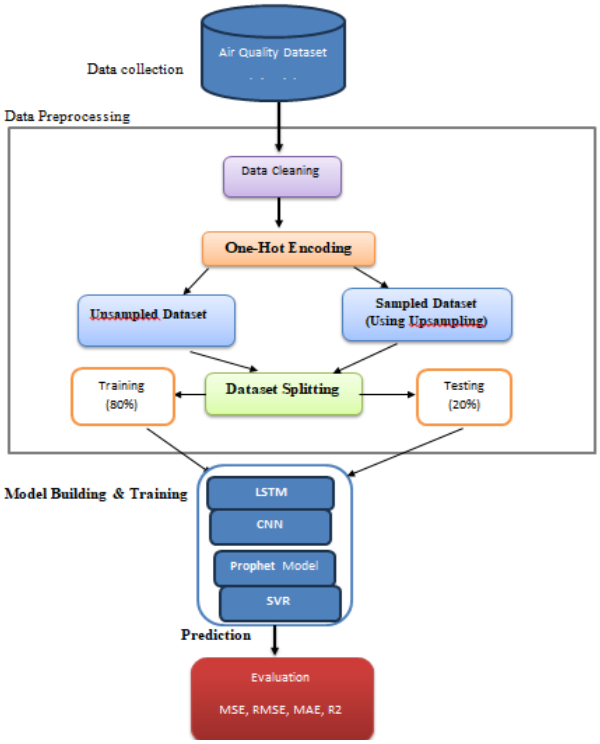

Figure 1: Block diagram of AQI proposed for Lagos State

Description of dataset

The dataset was obtained from the OpenWeatherMap API (https://openweathermap.org/api) which provides comprehensive and high-resolution air quality and meteorological data. The dataset covers a period from November 25th, 2020, to December 1st, 2023, and includes historical data from Apapa area of Lagos in Nigeria.

The data include several key variables such as the concentrations of different pollutants (e.g., PM2.5, PM10, NO2, SO2, CO, and O3), temperature, humidity, wind speed, and other meteorological factors as shown in Fig. 3. This rich dataset is essential for accurately modeling and predicting the Air Quality Index (AQI) as shown in Figs 2 and 3.

	2	12/21/2020 16:00	Apapa	Lagos State	South-West	403.88	0.26	5.14	25.75	1.49	6.06	21.11	3.64
0	1	11/25/2020 1:00	Apapa	Lagos State	South-West	287.06	0.0	1.54	35.05	0.62	2.98	8.50	1.38
1	1	11/25/2020 2:00	Apapa	Lagos State	South-West	283.72	0.0	1.34	36.12	0.70	3.13	8.32	1.55
2	1	11/25/2020 3:00	Apapa	Lagos State	South-West	293.73	0.0	1.48	35.76	0.86	3.47	8.82	1.93
3	1	11/25/2020 4:00	Apapa	Lagos State	South-West	343.80	0.0	2.31	34.69	1.43	5.04	11.18	3.14
4	2	11/25/2020 5:00	Apapa	Lagos State	South-West	547.41	0.0	4.88	30.76	2.92	12.33	21.21	6.21

Figure 2: Sample of AQI dataset for Lagos state

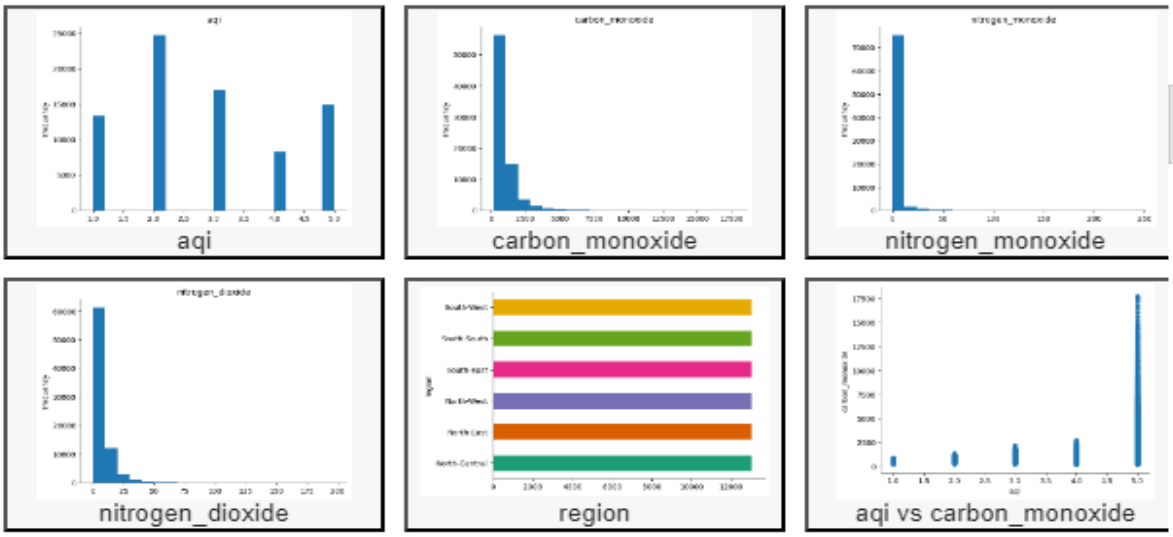

Figure 3: Visualization of the pollutants in the dataset

Data pre-processing

Pre-processing helps in changing the data into a better input data format that will be used in the model in order to improve the performance of the model. The study employed various pre-processing methods in order to increase the accuracy of the predictive model. The idea behind pre-processing is to ensure that the data is free from errors, inconsistencies, and missing values that could affect the model's performance.

Handling missing values

This provides that the dataset remains complete and reliable. Missing values were identified using methods like isnull() and isna(). The dropna() method was employed to remove rows with missing values.

Removing duplicates

Drop_duplicates() method was used to remove any duplicate rows, ensuring that each data point is unique.

Data type conversion

The date column was converted to datetime format for the AQI predictive model. So also columns were converted to appropriate data types (e.g., integers, floats, datetime) using astype() in pandas.

One-hot encoding

Categorical variables, such as AQI categories and state names, were encoded using one-hot encoding to be used in the models. In order to balance the dataset, SMOTE algorithm was used on the dataset which involves randomly duplicating examples from the minority class. The mathematical state as follows:

1. Identify the Minority Class: Let N_m be the number of instances in the minority class and N_M be the number of instances in the majority class.

2. Determine the Upsampling Factor: Calculate the upsampling factor as shown in Eqn. 2

$$f \text{ as } f = \frac{N_M}{N_m} \qquad (2)$$

3. Generate Synthetic Samples: For each instance in the minority class, duplicate it tf times (or a fraction thereof if partial duplication is used).

Model development

In order to come up with AQI predictive model for Lagos State, four models which are: LSTM, CNN and Facebook Prophet Model, SVR were examined in this study for the Lagos AQI predictive Model.

LSTM model

An LSTM network consists of three key gate components: the forget; the input; and the output. These gates govern the flow of information within the cell state, which help the network to run and update the cell state over long period.

1. Forget Gate: incharge of removing information from the cell state.

$$f_t = \sigma\left(w_f \cdot [h_{t\_1}, \varkappa_t] + b_f\right) \quad (3)$$

where f_t is the forget gate, w_f is the weight matrix, h_{t_1} is the previous hidden state, is the current input, b_f is the bias, and σ is the sigmoid function.

2. Input Gate: in control for the information stored in the cell state.

$$i_t = \sigma(w_i \cdot [h_{t\_1}, \varkappa_t] + b_i) \quad (4)$$

Where i_t is the input gate, w_i is the weight matrix, and b_i is the bias. Eqn.5

$$\tilde{c}_t = tanh(w_C \cdot [h_{t\_1}, \varkappa_t] + b_C) \quad (5)$$

Where $\tilde{c}_t$ is the candidate cell state, w_C is the weight matrix, and b_C is the bias.

In the cell State Update: old state and the new candidate state is joined together as shown in Eqn. 6.

$$Ct = ft * Ct - 1 + it \quad (6)$$

Where Ct is the updated cell state and * denotes element-wise multiplication.

3. Output Gate: Selects what information is produced based on the cell state. See Eqn. 7

$$O_t = \sigma(w_o \cdot [h_{t\_1}, \varkappa_t] + b_o) \quad (7)$$

Where O_t is the output gate, w_o is the weight matrix, and b_o is the bias. Eqn. 8

$$H_t = O_t * \tanh(Ct) \quad (8)$$

Where H_t is the new hidden state.

Prophet model

The mathematical Model for Facebook Prophet Model used for AQI prediction

The "Prophet Equation" fits, as shown in Eqn. 9: trends and seasonality.

$$y(t) = g(t) + s(t) + e(t) \quad (9)$$

where: g(t) refers to trend (changes over a long period of time); s(t) refers to seasonality (periodic or short-term changes); e(t) refers to the unconditional changes that is specific to a business or a person or a circumstance. It is also called the error term; y(t) is the forecast..

Parameter evaluation and model selection

In this work, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean absolute error(MAE) and R2 Score metrics were employed. These validation metrics are used to check whether DL regression models developed are perfect or deceptive.

MSE: is a metric used to measure the average squared difference between the actual values and the predicted values in a regression problem. The smaller the MSE, the closer the estimator is to the true values.

Mathematically as shown in Eqn. 10

$$MSE = \frac{1}{N}\sum_{i=1}^{N} \quad (yi - y^i)^2 \quad (10)$$

Where: N is the number of observations in the dataset; yi is the actual value of the dependent variable for the ith observation; y^i is the predicted value of the dependent variable for the ith observation.

RMSE: is a commonly used metric to evaluate the performance of a regression model. It measures the average of the squared differences between predicted and actual values, taking the square root of the result. Mathematically, it is defined as shown in Eqn. 11

$$\text{RMSE} = \sqrt{\left\{\frac{1}{n} sum_{\{i=1\}}^{\{n\}} (y_i - \hat{}\{y\}_i)^2\right\}} \quad (11)$$

Where: n is the number of observations in the dataset; $_i$ is the actual value of the dependent variable for the ith observation; y^i is the predicted (estimated) value of the dependent variable for the ith observation

MAE calculate errors between paired observations [16]. The greater the value of MAE , the bigger the error. The error is the difference between predicted value (ˆ Fi), actual value (Oi) and n is total number of observation. Eqn 12.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \quad \frac{(O_i - F_i)}{O_i} \; x\; 100 \quad (12)$$

Where Oi refers to the actual AQI and F_i refers to the predicted AQI and "n" refers to the total number of the prediction.

R^2 score calculates the proportion of variance in a case where independent variables define the dependent variable. The coefficient of determination (R^2 Score) is more correct and useful and is not affected from the interpretability issues [17]. The percentage of the dependent variable that can be forecast from the independent variable is referred to as the coefficient of determination, denoted as R^2. The equation for R^2 is given as Eqn 13.

$$R^2 = 1 - \frac{\sum_{i=1}^{D}(P_i - \hat{P}_i)^2}{\sum_{i=1}^{D}(P_i - \bar{P})^2} \quad (13)$$

Where $\hat{P}_i$ is the predicted ith value, P_i element is the actual ith value and $\bar{P}$ is mean of actual values and D is total number of observations.

Results and Discussion

In this study, the Lagos dataset mentioned above was cleaned so that the accuracy of the predictive model could be increased. Thereafter SMOTE was applied to the dataset to have a balanced version. The splitting of the dataset into training and testing datasets at the ratio of 80:20 respectively is shown in Fig. 4.

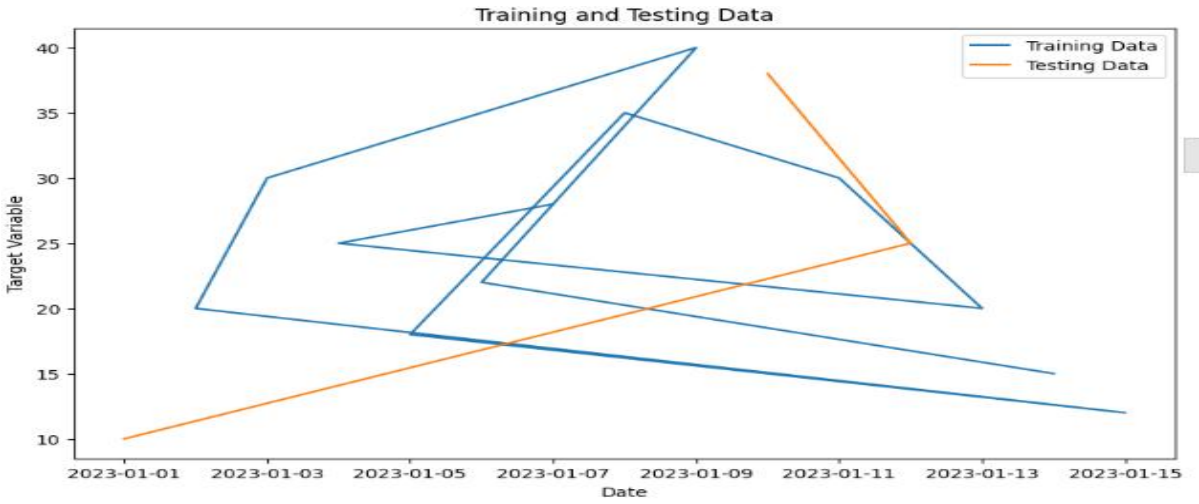

Figure 4: Training and testing datasets

LSTM

The LSTM model showed the state-of-the-art results in AQI prediction [18]. The LSTM model developed and trained with 256 cell along with Relu activation function and 2 Dense layer with one neuron along with linear activation function. The comparison of predicted values using LSTM model and actual values is represented in Figs 5 and 6.

	Original AQI	Predicted AQI
50196	3	2.916026
44476	3	3.211745
67604	5	4.989516
11153	1	1.199193
68276	5	4.983315
21354	1	1.186206
5904	2	1.640822
72690	5	4.293113
75943	5	4.978881
17749	4	3.527476
39168	3	3.019307
45031	3	3.097824
44796	3	2.854037
68919	5	4.267480
58808	4	3.882603
24439	5	4.951043
40230	3	3.425989
41818	3	3.610950
21356	1	1.463684
3789	1	1.065672

Figure 5: Predicted AQI

Table 2: Lagos AQI categories and its effects

S/No	Original AQI	Predicted AQI	Description
1	3	2.9	Low
2	3	3.2	Low
3	5	4.9	Medium
4	1	1.1	Low
5	5	4.9	Medium
6	1	1.2	Low
7	2	1.6	Low
8	5	4.2	Medium
9	5	4.9	Medium
10	4	3.5	Medium
11	3	3.0	Low
12	3	3.1	Low
13	3	2.9	Low
14	5	4.3	Medium
15	4	3.9	Medium
16	5	4.9	Medium
17	3	3.4	Low
18	3	3.6	Low
19	1	1.4	Medium
20	1	1.1	Low

The result of the prediction showed that the pollution is within the range of low and medium as seen in Table 2. The *p*-value is less than the significance level of 0.05. This confirmed that the series is not a stationary time series. The first order differentiation time series is presented in Figs 7 and 8 for AQI of Lagos, Autocorrelation function and partial autocorrelation function plot is seasonality in time series as seen in Fig. 9. The first-order difference series is analyzed (Annual seasonality lag =5 period). The comparative analysis of Augmented Dickey-Fuller (ADF) tests of first order difference series with 5 period lag is provided in Table 3. The results depict that the time series after the first order differentiation (lag=5) is not purely random.

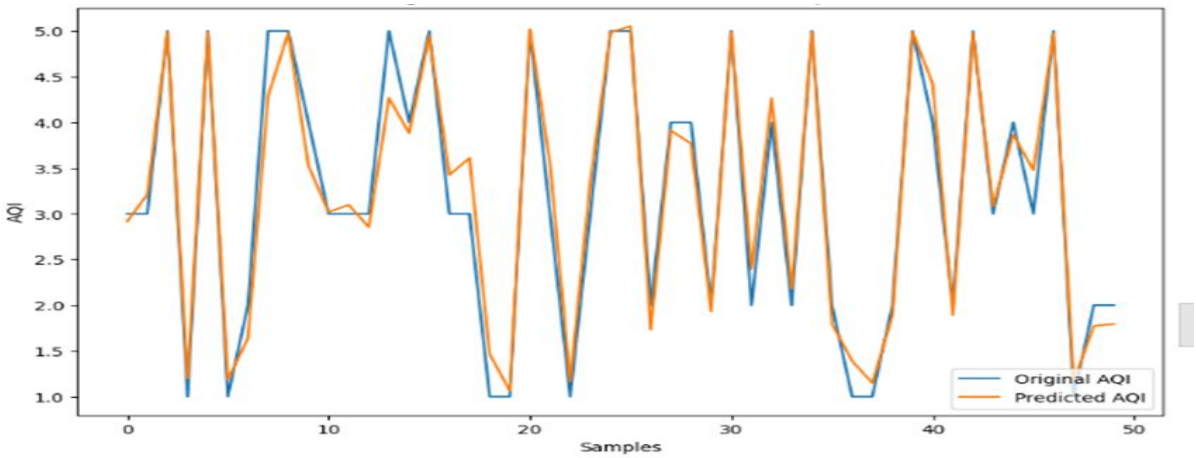

Figure 6: original vs predicted AQI (first 50 samples)

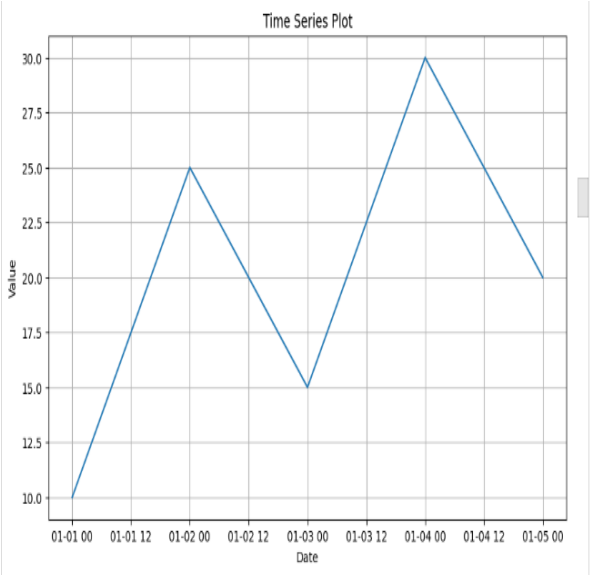

Figure 7: Time series

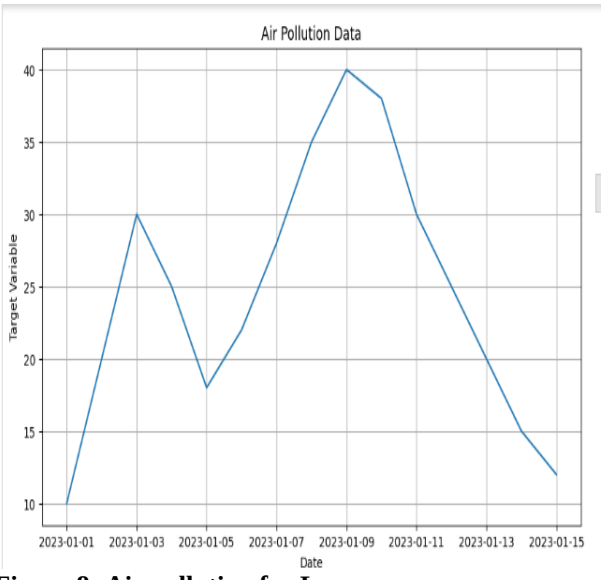

Figure 8: Air pollution for Lagos

Table 3: ADF test results on Lagos AQI time series.

Test Statistic	-5.753965
P-value	0.0004270
Critical Value	2.2281388

Table 4: Validation metrics

S/N	Metric	Value
1	MSE	0. 062
2	RMSE	0.249
3	MAE	0. 149
4	R-squared:	0. 968

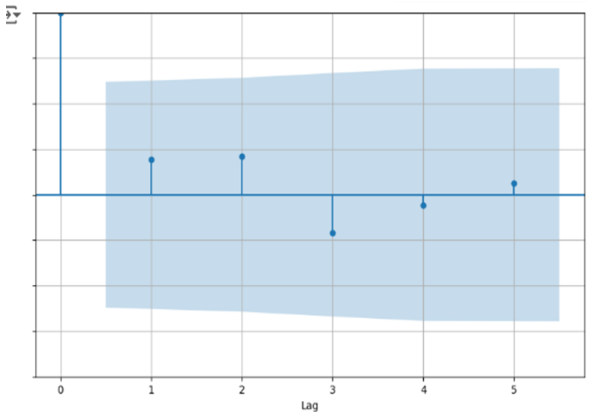

Figure 9: Autocorrelation function

CNN Model

CNN is one of the deep learning models suitable for processing data that have grid patterns, such as images, which is inspired by the organization of animal visual cortex Fukushima [19]. It is designed to automatically and adaptively learn spatial hierarchies of features, from low- to high-level patterns. CNN has three layer which are convolutional layers, pooling layers and dense layer.

Convolutional layers: these layers apply convolution operations to the input. A convolution operation involves sliding a filter (or kernel) over the input image

to create feature maps. Mathematically, this can be represented as shown in Eqn.14

$$(f \times g)(i, j) = \sum \sum f(m, n)\, g(i - m, j - n) \quad (14)$$

Where ***f*** is the input feature map and **g** is the filter

Activation functions: after the convolution operation, an activation function (commonly ReLU) is applied to introduce non-linearity as seen in Eqn. 15

$$\mathrm{ReLU}_{(x)} = \max(0, x) \quad (15)$$

The CNN model applied to predict the AQI is validated using some metric as shown in Table 5 in order to ascertain the effectiveness of the AQI predictive model developed. The MSE shows that model developed learn better as the epoch increases as shown in Figs 10 and 11. The accuracy also improves as the epoch increases as shown in Fig. 10.

Table 5: Analysis of AQI using CNN Model

S/N	Metric	Value
1	MSE	0.110
2	RMSE	0.332
3	MAE	0.231
4	R-squared:	0.231

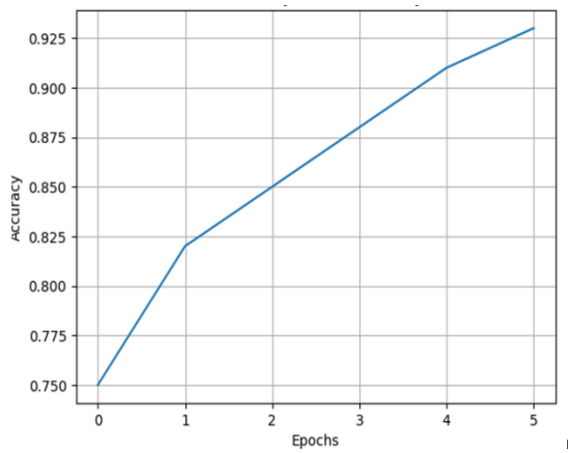

Figure 10: Air quality accuracy

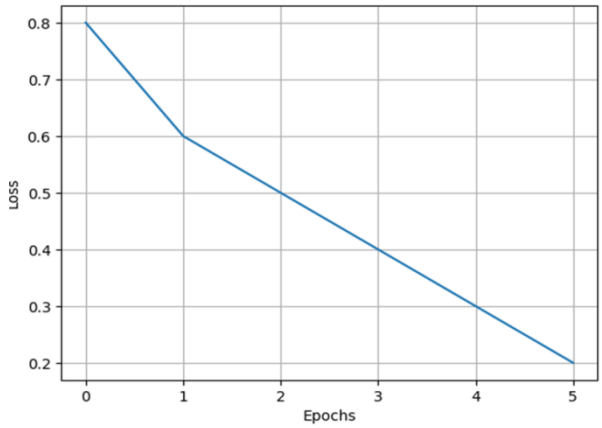

Figure 11: Training loss

Prophet Model

Facebook's Prophet Technique is built on an addictive-regressive technique. The following elements make up the fb prophet model as seen in Eqn 16:

$$(p) = (p) + (p) + \in p \quad (16)$$

Here, g(p) is the trend of the time series data, s(p) denotes its seasonal pattern and p denotes the model error. The model is created using Google Colaboratory and it requires inputs parameters and the target variable to be predicted which is AQI as seen in Fig 12.

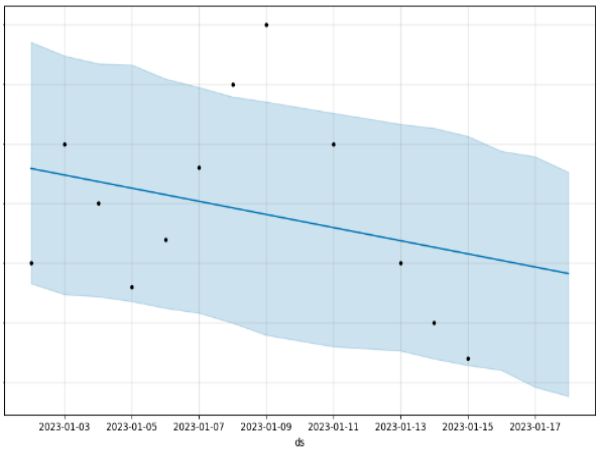

Figure 12 FB Prophet Model

The Prophet Model forecast plot provides a visual representation of the forecast values, historical data, and uncertainty intervals. It helps to understand the predicted trends, seasonality, and potential deviations from the forecast. Black Dots represents the historical data points used to train the model as seen Fig. 12. Blue Line represents the central tendency of the forecast showing the predicted values generated by the Prophet model while the Light Blue Shaded Region represents the uncertainty intervals or confidence. The FB Prophet model predicted the AQI on monthly basis, daily basis and hourly basis as shown in Figs 13, 14 & 15. Table 6 revealed the validated metrics used to validate the model.

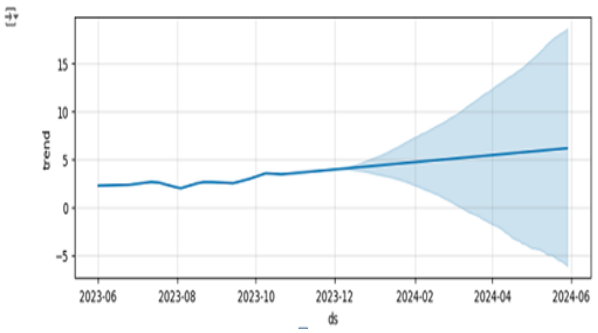

Figure 13: FB Prophet Model prediction on monthly basis

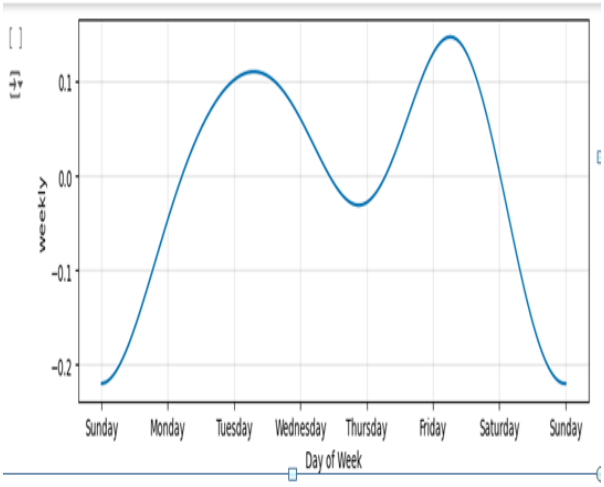

Figure 14: FB Model prediction on daily basis

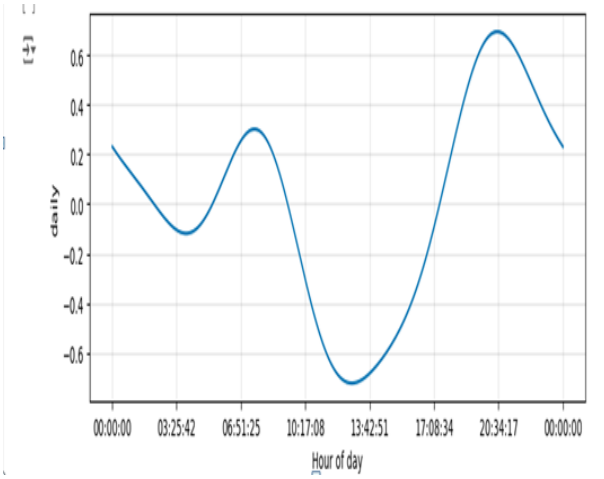

Figure 15: FB Model prediction on hourly basis

Table 6: Analysis of results of Prophet Model for AQI prediction

S/N	Metric	Value
1	MSE	2.258
2	RMSE	1.502
3	MAE	1.289
4	R2	-0.377

Table 7: Analysis of results of SVR for AQI prediction

S/N	Metric	Value
1	MSE	0.097
2	RMSE	0.312
3	MAE	0.215
4	R-squared:	0.949

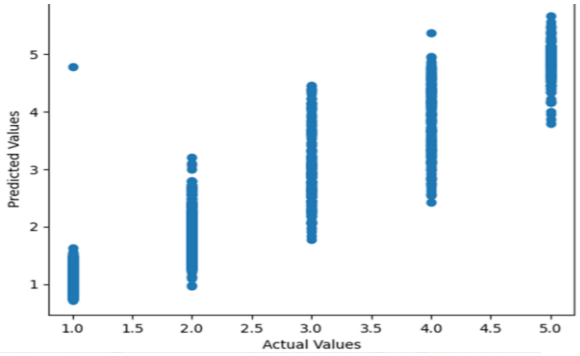

Figure 16: SVR Model performances

SVR Model

SVR model is used to transform linearly separable data into high dimensional space with the help of kernel functions [20]. In order to analyze the behavior of the data and performance of the SVR kernel function on these data, the outcome of the four different validation metrics MSE, RMSE, MAE, R^2 of the SVR model is presented in Table 7. The predicted value and actual value of the AQI is shown in Fig. 16.

Table 9 records the result of performance metrics used for the Lagos state with balanced dataset (i.e.,) using the SMOTE for all four algorithms such as LSTM, CNN, Prophet Model and SVR. The LSTM model gives the best result in comparison to the CNN, Prophet Model and SVR.

Table 9 Comparison of the four models developed

Predictive Model	MSE	RMSE	MAE	R2
LSTM Model	0.062	0.249	0.149	0.968
CNN	0.1102	0.332	0.231	0.2310
Prophet	2.258	1.502	1.289	-0.377
SVR Model	0.097	0.312	0.215	0.949

In Fig. 17, comparison between MSE of LSTM, CNN, Prophet Model and SVR with balanced dataset using the SMOTE algorithm is shown. It depicts that LSTM has lowest MSE when compare with other model such as CNN, Prophet and SVR. In Fig. 18, the comparison between RMSE of LSTM, CNN, Prophet Model and SVR with balanced dataset using the SMOTE algorithm is shown. It depicts that LSTM has lowest RMSE when compare with other model such as CNN, Prophet and SVR.

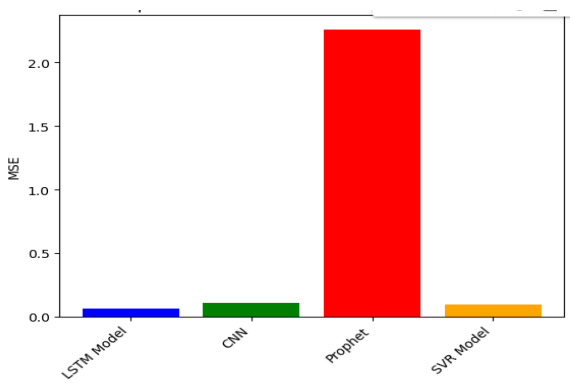

Figure 17: Comparison of MSE Metric

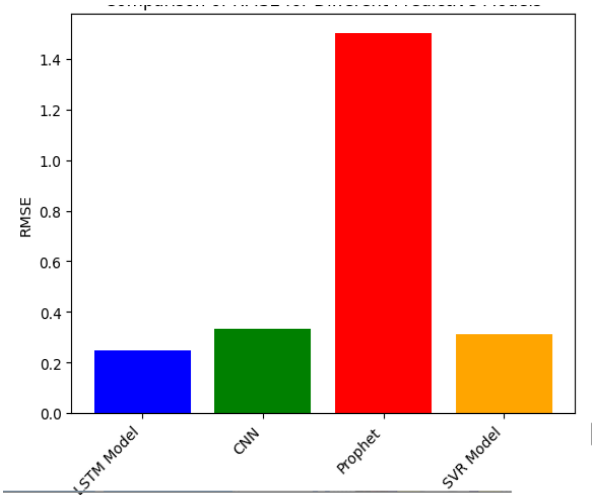

Figure 18: Comparison of RMSE Metric

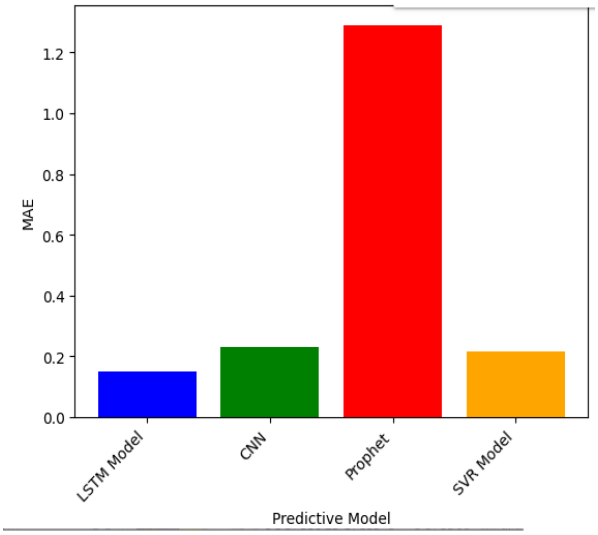

Figure 19: Comparison of MAE Metric

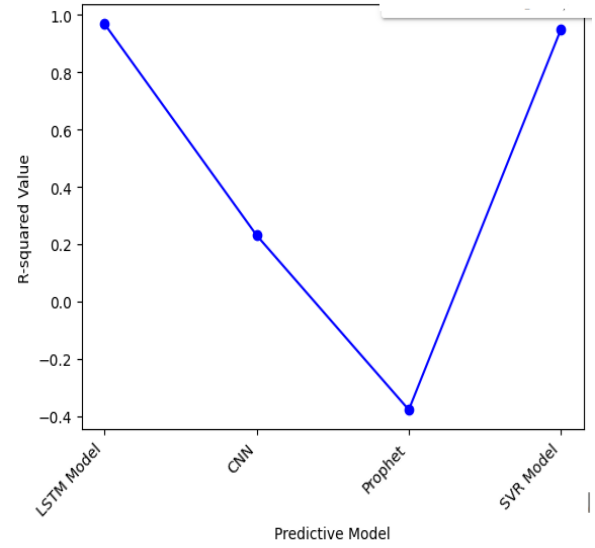

Figure 20: Comparison of R2 Metric

In Fig. 19, the comparison between MAE of LSTM, CNN, Prophet Model and SVR with balanced dataset using the SMOTE algorithm is shown. It depicts that LSTM has lowest MAE when compare with other model such as CNN, Prophet and SVR. In Fig.re 20, the comparison between R-SQUARE of LSTM, CNN, Prophet Model and SVR with balanced dataset using the SMOTE algorithm is shown. It depicts that LSTM has the highest R-SQUARE, when compare with other model such as CNN, Prophet and SVR.

LSTM Model has the lowest MSE (0.062), indicating it is the best performing model among the four.(LSTM Model, CNN, Prophet ,SVR Model).So also, LSTM Model has the lowest RMSE, suggesting it has the smallest average error in its predictions compared to the other models. LSTM Model has the lowest MAE (0.1490), indicating it is the best performing model among the four. Mean Absolute Error (MAE) measures the average absolute difference between the predicted and actual values. LSTM Model perform better with an R2 of 0.968 compared to other models. Therefore, LSTM outperform all other models developed for LAGOS AQI. The results of the model developed in this study is compared with the results presented in Nilesh and Safvan [21] on prediction of AQI using machine learning for Ahmedabad city where the model achieved R2= 0.951236947, MSE= 656.6232392 and RMSE = 25.62466076. The result of the proposed LSTM model is better in terms of the evaluation parameters: MSE= 0. 062; RMSE =0.249; MAE =0. 149 and R2= 0. 968.

Conclusion

The paper come up with efficient model for the AQI prediction using the air quality dataset for Lagos state obtained from the OpenWeatherMap API. Various pre-processing methods were used for improved data representation such as outlier removal, data normalization, feature selection and missing value management. Four different machine learning models: LSTM; CNN; Prophet Model and SVR have been investigated in this study to produce an effective predictive model. When predicting the AQI data for Lagos State the LSTM outperformed other models such as CNN, Prophet Model and SVR on the validating metrics known as MAE, MSE R^2 score and RMSE. The model obtained could be subjected to further research in other geographic regions, such as Delta State or other state by state level analysis which could be expanded to forecast other pollution indices at different levels. More feature engineering techniques can be employed for accuracy comparison.

Conflict of interest: No conflict of interest.

Acknowledgment: We hereby acknowledge and thank all authors cited in this paper.

References

[1] Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D. & Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525(7569), 367-371.

[2] Seinfeld, J. H. & Pandis, S. N. (2016). Atmospheric Chemistry and Physics: From Air Pollution to Climate Change. (3rd ed.).

[3] U.S. Environmental Protection Agency (2020). Ozone pollution. EPA.

[4] Kampa, M. & Castanas, E. (2008). Human health effects of air pollution.*Environmental Pollution*, 151(2), 362–367.

[5] World Health Organization (WHO) (2016). Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease. World Health Organization. https://www.who.int/publications/i/item/9789241511353/

[6] European Environment Agency (EEA) (2021). Air Quality Index. European Environment Agency.

[7] Health Effects Institute (2019). State of Global Air 2019. Health Effects Institute.

[8] Akinyemi, S. O. & Adedeji, O. T. (2019). Impact of air pollution on public health in Nigeria: A critical review. *Environmental Science and Pollution Research*, 26(3), 2501-2510.

[9] Yale Center for Environmental Law & Policy. (2018). Environmental Performance Index 2018. Yale University. Retrieved from https://epi.yale.edu/

[10] Yakubu, O. O. (2017). A review of the effects of air pollution on human health in Nigeria. *African Journal of Environmental Science and Technology*, 11(3), 112-121.

[11] Ede, J. A. & Edokpa, F. (2017). Assessment of air pollution and human health risks in selected urban areas of Nigeria. *Journal of Environmental Protection*, 8(8), 1056-1066.

[12] Akinfolarin, O.M., Boisa, N. & Obunwo, C.C. (2017). Assessment of particulate matter-based air quality index in Port Harcourt, Nigeria. *Journal of Environmental Analytical Chemistry*, 4, 1-4. https://doi.org/10.4172/2380-2391

[13] Srinivasa Gupta , Yashvi Mohta, Khyati Heda, Raahil Armaan, Valarmathi, B. & Arulkumaran, G. (2023). Predicting urban air quality using machine learning: A case study from India. *Atmospheric Environment*, 271, 118504.

[14] Jingyang Wang, Xiaolei Li, Lukai Jin, Jiazheng Li, Qiuhong Sun & Haiyao Wang (2022). An air quality index prediction model based on CNN-ILSTM. *Scientific Reports*, 12, 8373. https://doi.org/10.1038/s41598-022-12355-6

[15] Alaba, O. B., Ogunsanwo, G. O., Olayinka, O. R., Taiwo, E. O. & Abass, O. A. (2021). Effective analysis of air pollution using decision trees, Naive Bayes and ZeroR Classiffier. *Journal of Science and Information Technology*.

[16] Sammut, Claude, Webb, G. I. (2010). Mean squared error. In: Sammut, Claude & Webb G. I. (Ed.), *Encyclopedia of Machine Learning*, Springer US, Boston, MA, p. 653, 10.1007/978-0-387-30164-8_528

[17] Chicco, D., Warrens M.J & Jurman G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation Peer. *J. Comput. Sci*., 7, Article e623

[18] Seng D., Zhang Q., Zhang X., Chen G. & Chen X. (2021). Spatiotemporal prediction of air quality based on LSTM neural network Alex. *Eng. J*., 60(2).

[19] Fukushima, K. (1980). Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.*, 36, 193–202. https://www.epa.gov/ozone-pollution/

[20] Dun, M., Xu, Z., Chen, Y. & Wu, L. (2020). Short-term air quality prediction based on fractional grey linear regression and support vector machine. *Math. Probl. Eng.*, 2020 (2020), pp. 1-13. 10.1155/2020/8914501

[21] Nilesh, A. & Safvan, S. (2023). The role of remote sensing in air quality monitoring in urban environments. *International Journal of Remote Sensing*, 44(4), 1290-1309.

Citing this Article

Ogunsanwo, G. O., Odulaja, P. T., Omotunde, A. A. & Solanke, O. O. (2025). Air quality index prediction using deep learning for Lagos State in Nigeria. *Lafia Journal of Scientific and Industrial Research, 3*(1), 98 – 107. https://doi.org/10.62050/ljsir2025.v3n1.450