A Machine Learning Model for Predicting Carbon Emission

Emmanuel Bamidele Ajulo[1*], Raphael Olufemi Akinyede[2], Shukurat Adeteju Bello[3]

[1]Department of Computer Science, Federal University of Technology, Akure, Ondo State, Nigeria.

[2]Department of Information Systems, Federal University of Technology, Akure, Ondo State, Nigeria.

[3]Department of Computer Science, Caleb University, Imota-Ikorodu, Lagos State, Nigeria.

***Corresponding Author**

DOI: https://doi.org/10.51584/IJRIAS.2026.11010028

Received: 24 December 2025; Accepted: 29 December 2025; Published: 29 January 2026

ABSTRACT

Air pollution impacts human health in various ways, including by depleting the ozone layer. This study aimed to utilise available data to develop a machine-learning model that predicts carbon emissions. The dataset was processed, converted to a time series, and split into training and test sets at a 70:30 ratio. The Long Short-Term Memory (LSTM) and Autoregressive Integrated Moving Average (ARIMA) models were employed to develop the model. Root Mean Squared Error (RMSE) metrics were used to evaluate the results. The findings indicate that applying the LSTM model to a large dataset with a high number of epochs yields better accuracy than using ARIMA on the same dataset. The LSTM achieved a lower RMSE of 0.0440 and better predicted carbon emissions than ARIMA. The system developed is recommended for countries, organisations, and agencies to monitor carbon-related air pollution.

Keywords: Machine learning, Dataset, Air pollution, Carbon, Long Short-Term Memory, Autoregressive Integrated Moving Average.

INTRODUCTION

Air pollution refers to the harmful effects of sources that contribute to atmospheric pollution and the deterioration of ecosystems (WHO, 2019). It consists of many types of pollutants, including materials in solid, liquid, and gaseous forms, caused by both human activities and natural phenomena (Ghorani-Azam et al., 2016). In recent years, air monitoring has garnered increased interest across various fields. Many efforts have been focused primarily on applications related to citizens' health and safety (Molinara et al., 2019). Pollution sources range from small-scale emissions, such as cigarette smoke and volcanic activity, to large-scale emissions from automobile engines and industrial processes (Ghorani-Azam et al., 2016).

According to WHO (2019), 4.2 million people are estimated to die each year due to diseases caused by air pollution, such as heart disease and stroke. Air pollution has various health effects. The health of susceptible individuals can be affected even on days with low air pollution. Short-term exposure to air pollutants is closely linked to COPD (Chronic Obstructive Pulmonary Disease), cough, shortness of breath, wheezing, asthma, respiratory diseases, and higher hospitalisation rates (a measure of morbidity). The long-term effects associated with air pollution include chronic asthma, pulmonary insufficiency, cardiovascular diseases, and cardiovascular mortality. Air pollution also negatively impacts the climate and the environment as a whole. It diminishes the quality of our Earth. Pollutants such as black carbon, methane, tropospheric ozone, and aerosols influence the amount of incoming sunlight.

As a result, the Earth's temperature is rising, leading to the melting of ice, icebergs, and glaciers (Molinara et al., 2019). With this in mind, it is clearly urgent to improve air quality and implement measures to protect the environment. Air quality prediction can be a valuable investment at multiple levels—individual, community, national, and global. Accurate predictions help people plan, reduce the impacts of harmful air pollutants on

health and associated costs, and contribute to a cleaner, healthier environment (Sanjeev, 2021). One of the greatest threats to air quality is carbon-related air pollution, as well as other pollutants released into the atmosphere. This is why the air quality parameter considered in the proposed system is carbon. It is classified as a smothering gas that can reduce the availability of oxygen to tissues. It is a toxic substance that can harm the body.

Long- and short-term exposure to airborne toxicants has different toxicological effects on humans, including respiratory and cardiovascular diseases, neuropsychiatric issues, eye irritation, skin conditions, and chronic diseases such as cancer. Several reports have shown a direct link between poor air quality exposure and increasing rates of morbidity and mortality, mainly due to cardiovascular and respiratory diseases. Air pollution is recognised as a significant environmental risk factor in the development and progression of conditions such as asthma, lung cancer, ventricular hypertrophy, Alzheimer's and Parkinson's diseases, psychological issues, autism, retinopathy, foetal growth problems, and low birth weight.

Air quality prediction can be a valuable investment on multiple levels – individual, community, national, and global. Accurate forecasts help people plan, reducing the impact of harmful air pollutants on health and lowering related costs, while fostering a cleaner, healthier environment (Sanjeev, 2021). The World Health Organisation (WHO) reports on six major air pollutants: particle pollution, ground-level ozone, carbon, sulphur oxides, nitrogen oxides, and lead. Air pollution can severely affect all components of the environment, including groundwater, soil, and air. Moreover, it poses a significant threat to living organisms (Molinara et al., 2019). Pollution sources range from small-scale origins like cigarette smoke and natural events such as volcanic activity to significant emissions from vehicle engines and industrial processes. Nevertheless, researchers have employed various models, such as Random Forest and Support Vector Machine, to address carbon emissions (Sanjeev, 2021).

In this study, the primary focus is on carbon emissions, which are the most severe in our ecosystems; these include excessive emissions from unchecked industrial activities, the proximity of industrial areas to residential zones, imported second-hand vehicles, numerous trucks and trailers on the roads, and old, poorly maintained transport vehicles operating without proper checks or measures to remove them from the streets. The approach of this study is to develop machine learning systems that learn and improve performance over time based on data to predict carbon emissions. Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) algorithms were employed, and their results were compared and evaluated.

LITERATURE REVIEW

Castelli et al. (2020) employed support vector regression (SVR) to forecast pollutant and particulate levels and to predict the air quality index (AQI) for California, United States. (Bhardwaj and Sharma, 2021) used the Support Vector Regression (SVR) model to forecast levels of pollutants (NO2, SO2, PM2.5, and PM10) and the AQI, using publicly available data for New Delhi. (Liu et al., 2020) developed a quantitative remote sensing-based algorithm for air pollution monitoring using artificial intelligence. They proposed an artificial intelligence-based quantitative monitoring algorithm for air pollution, and the experimental results show that it achieves high monitoring efficiency, a wide monitoring range, and high accuracy.

Bhardwaj and Sharma (2021) developed an innovative solution to reduce respiratory problems caused by indoor air pollution. Their research focused on testing whether placing certain plants indoors can reduce indoor air pollution, specifically particulate matter (PM), using principles of biomedical engineering and machine learning. The study was limited to particulate matter (fine dust particles suspended in the air) pollution in indoor spaces, and challenges with real-time PM level data collection using sensors were encountered. Heydari et al. (2021) proposed a new hybrid intelligent model based on long short-term memory (LSTM) and the multi-verse optimisation algorithm (MVO) to predict and analyse air pollution, specifically nitrogen dioxide (NO2) and sulphur dioxide (SO2), from Combined Cycle Power Plants. To assess the performance of the proposed model, it was applied to real data from a Combined Cycle Power Plant in Kerman, Iran. Meivel et al. (2021) introduced a system that provides an efficient method for detecting air quality. They developed a model that uses sensors to measure environmental gas levels, focusing on carbon dioxide, dust, and methane. The sensor signals are fed

into a microcontroller, with threshold levels fixed for each gas. The sensed data is stored in the cloud, and when gas levels exceed the thresholds, an alarm alerts relevant personnel to notify employees.

A study by Wood (2022) utilises air quality data from a compiled daily-averaged dataset for six air pollutants in Dallas County for the period 2015 to 2020. The daily data is derived from the United States National Air Quality System (AQS) database maintained by the Environmental Protection Agency. Babu Saheer et al. (2022) sought to address the challenge of climate change by leveraging the availability of big data across different domains, such as pollutant concentration, urban traffic, aerial imagery of terrains and vegetation, and weather conditions, which can assist in understanding the interactions between these factors and in constructing a reliable air quality prediction model. They propose a novel, cost-effective, and efficient air quality modelling framework that incorporates all the factors mentioned above and employs state-of-the-art artificial intelligence techniques. The framework also features a novel deep learning-based vegetation detection system using aerial imagery. The pilot study, conducted in the UK city of Cambridge, uses the proposed framework to explore predictive models spanning statistical methods, machine learning, and deep recurrent neural networks. This study aims to apply artificial intelligence to predict carbon emissions using global carbon data sourced from 'carbonmonitor.org'.

METHOD

This work aims to develop a machine learning model to predict the amount of carbon emitted at a specific time using data from the 'carbon monitor' web platform and predictive analysis.

The Proposed Model

Long Short-Term Memory (LSTM) and ARIMA are employed to develop models, and their performance is compared to identify the most accurate prediction. The architecture of the predictive carbon emission system is depicted in Figure 1. The system is also divided into five modules, including: Carbon Monitor Dataset, Data Preprocessing, Model Splitting (Training and Testing Phase), Model Evaluation, and Results.

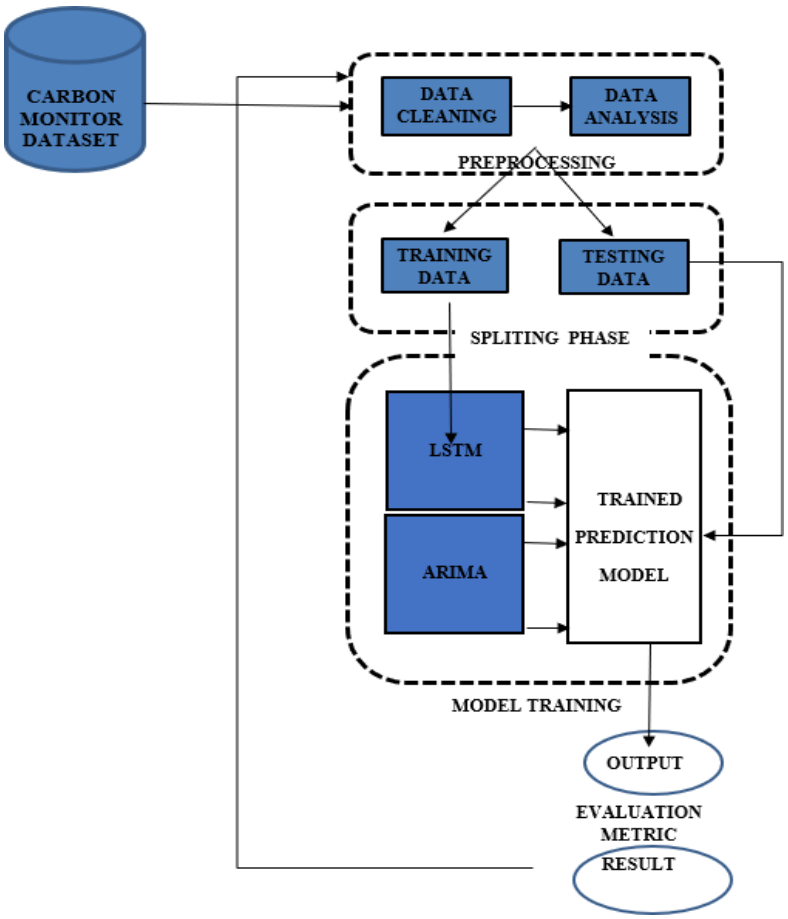

Figure 1: Architecture of the System.

Data Collection Approach

The dataset used in this study includes information on the value of a carbon pollutant and was downloaded from Carbon Monitor ('https://carbonmonitor.org/'), an international initiative that provides, for the first time, regularly updated, science-based estimates of daily carbon emissions. The data reveal the decline and subsequent rebound of emissions during the COVID-19 pandemic and the post-COVID-19 period. The dataset initially contains five attributes: country, date, sector (the sector responsible for generating the carbon value), value (the amount of carbon at a specific time), and timestamp, comprising 99,624 rows.

Data Preprocessing

This step cleans dirty/noisy data, extracts and merges data from different sources, and then transforms it into a standard format. The main dataset was converted into a time series dataset by selecting the 'date' and 'carbon value' variables as the features of the time series to be fed into the ARIMA and LSTM models.

Data Splitting

The time series dataset was split into two subsets: a training set and a test set, with 70% of each used for training and the remaining 30% for testing the model's accuracy. During training, the models 'learn' from the dataset and perform specific actions; during testing, their effectiveness and accuracy are assessed. Any errors or miscalculations are fed back into the training phase for the classifier to relearn from the data.

The Models

The Autoregressive Integrated Moving Average (ARIMA) is a univariate model designed to identify patterns using a simple algorithm. It works by first splitting a sequence of n time steps into a historical window and a prediction window. The size of the historical window depends on the ARIMA model's parameters. Before inputting the data into the model, the historical window undergoes differencing to make the time series more stationary. The ARIMA model then computes results based on the differenced data and compares them with the actual values.

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that can remember values from earlier time steps for future use. The three fundamental requirements of a recurrent neural network are as follows: that the system can store information for an arbitrary period; that the system is resistant to noise (i.e., fluctuations of the inputs that are random or irrelevant to predicting a correct output); and that the system parameters are trainable (within a reasonable time). The model has been built in Python.

Performance Evaluation Metrics

The model's performance is evaluated using Root Mean Squared Error (RMSE). The purpose of using RMSE is to represent the standard deviation of errors when a prediction is made on a data set, thereby assessing the model's accuracy. The Root Mean Square Error (RMSE) is a crucial metric for evaluating regression model performance. A lower RMSE indicates a better fit to the data, with predictions closer to the observed values. A value of 0 would represent a perfect, error-free model, which is highly unlikely in practice. It is calculated as:

$$\text{RMSE} = \sqrt{{}^{1}/_{N} * \sum_{t=1}^{N} (Yi - yi)^2} \quad (1)$$

RESULTS AND DISCUSSION

The system interface shown in Figures 2 and 3 displays the RMSE for both models as a function of selected input parameters. The model was fed a dataset containing 1000 attributes, the effect of SGD and Adam was checked as optimizers on the result and specified epoch as one; i.e the number of epochs is a hyperparameter that defines the number of times the learning algorithm will work through the entire training dataset, batch size

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume XI Issue I January 2026

was also specified to be one, this means the dataset will pass through the model as a whole without being fragmented. Based on several test iterations and the nature of the dataset, the epoch and batch sizes have been standardised to 1, 50, and 100. The summary of the result is shown in Table 1.

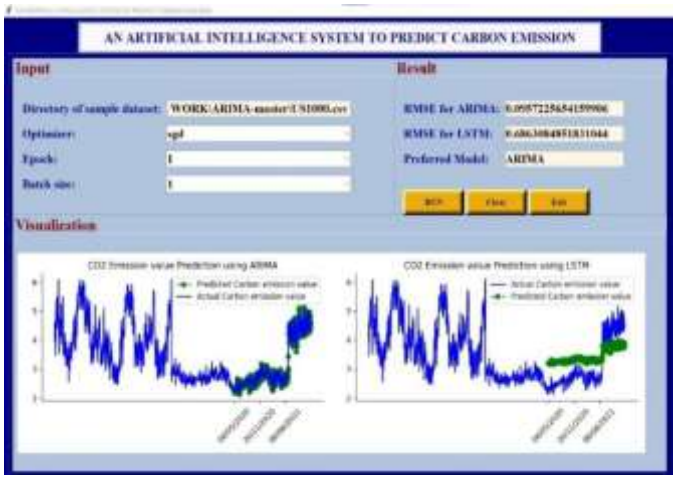

Figure 2: RMSE of ARIMA

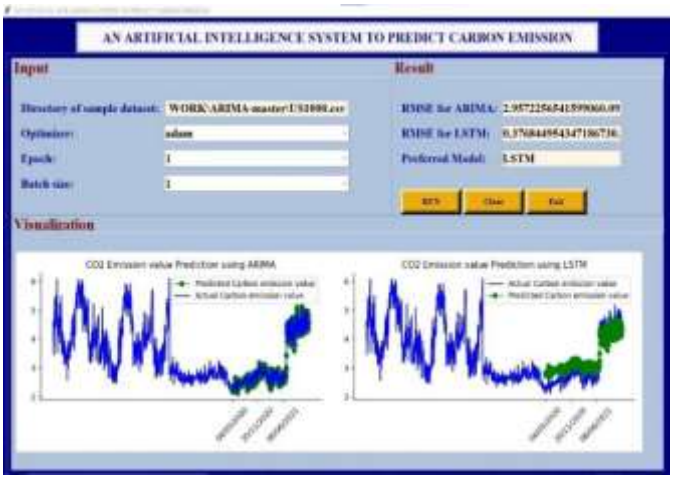

Figure 3: RMSE of LSTM

Table1: Summary of First Result

		ARIMA (RMSE)	LSTM (RMSE)			ARIMA (RSME)	LSTM (RSME)
Dataset	1000	0.0957	0.6863	Dataset	1000	2.9572	0.3768
Optimizer	SDG			Optimizer	Adam		
Epoch	1			Epotch	1		
Batch Size	1			Batch Size	1		

Using the Stochastic Gradient Descent (SGD) optimiser, ARIMA achieved a lower RMSE of 0.0957 and performed best based on the selected inputs. However, with the Adaptive Moment Estimation (Adam) optimiser, the LSTM model achieved a lower RMSE of 0.3768, which is better than ARIMA's 2.9572.

In Figures 4, 5, and 6, the effect of using a larger dataset (2000 attributes) on the model's performance was investigated. SGD was used to optimise all instances in the Figures, with epochs of 1, 50, and 100 in the three cases. Figure 7 shows that the model was randomly tested with a very small amount of data (100 attributes), and that Arima performed better even when the Epoch was very high. The dataset was found to be a key indicator of the model's performance, and the LSTM model does not perform optimally with a small dataset. The summary of the result is shown in Table 2.

Figure 4: Effect of SGD optimisation with epochs of 50

Figure 5: Effect of SGD optimisation with epochs of 1

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume XI Issue I January 2026

Figure 6: Effect of SGD optimisation with epochs of 100.

Figure 7: shows the model randomly tested with small dataset of 100 attributes,

Table 2: Summary of Second Result

		ARIMA RMSE	LSTM RMSE		ARIMA RMSE	LSTM RMSE		ARIMA RMSE	LSTM RMSE
Dataset	2000	0.1782	0.1321		0.1782	0.0440		0.1782	0.0452
Optimizer	SGD								
Epoch	1			50			100		
Batchsize	1								

The LSTM Model with epoch 50 was the most accurate with the lowest RMSE score of 0.440. It was also observed that, despite the change in the Epoch number specified, the RMSE of the ARIMA model remained the same, mainly because it is not a learning model but a statistical tool, and no matter how many times the model is trained on the dataset, the result remains the same. LSTM, on the other hand, is a learning algorithm, and it is expected to display better results with increased epochs.

The dataset size was found to affect RMSE, model performance, and the visualisation of predictions; the larger the dataset, the better. The performance of LSTM models improves with more data and training epochs because they are complex deep learning models designed to capture intricate, non-linear patterns and long-term dependencies, which require substantial data to learn effectively and avoid overfitting. In contrast, ARIMA is a simpler, traditional statistical model that assumes linear relationships and has a fixed structure, so its performance does not significantly benefit from excessive data or epochs beyond what is needed to estimate its few parameters. However, the larger the dataset, the slower the computation time. It was also observed that as the epoch size increased, the LSTM model performed better and increased in accuracy. Applying an LSTM model to a large dataset with many epochs will achieve better accuracy than using ARIMA on the same dataset. LSTM has the lowest RMSE of 0.0440 and will better predict carbon emissions. RMSE provides a standardised measure for comparing models on the same dataset. The model with the lowest RMSE is generally considered the most accurate. Interpreting an RMSE value requires context. An RMSE of 0.0440 is excellent for a variable with 99,624 rows used in this study.

The results of this study indicate that LSTM networks achieve superior predictive accuracy compared to ARIMA, particularly as data volume and the number of training iterations increase. These findings are consistent with prior research showing that recurrent neural networks outperform linear models when applied to complex, non-linear temporal processes.

Limitation And Future Work

This study has some limitations that should be acknowledged. The dataset used in this analysis is limited in scope, covering emissions data for 6 months. As a result, the findings may not fully generalise to other time periods. In this case, model performance can be prone to variance and over-fitting. It is expected that future work would consider using emission data covering a considerable period of time. The modelling approach used for this study is also univariate, relying solely on historical emissions values. Potentially influential exogenous factors, such as meteorological conditions, economic activity, population growth, and policy interventions, were not incorporated. Including such variables in a multivariate framework could improve predictive performance and provide deeper insights into the drivers of emissions trends. Computational constraints influenced confident modelling choices, including the selection of training configurations such as batch size and number of epochs for the LSTM model. While these settings were sufficient to demonstrate comparative performance trends between models, more extensive hyperparameter tuning and longer training schedules may yield further improvements in predictive accuracy. Future work could address these limitations by expanding the dataset, incorporating additional explanatory variables, and conducting more comprehensive model optimisation.

CONCLUSION

The main aim of this work was to predict carbon emissions, which significantly contribute to various environmental and climate issues worldwide. The study found that accurately forecasting carbon emission using ARIMA and LSTM yields different results depending on the input variables, with the LSTM model performing better. This study is a valuable investment on multiple levels; individuals, communities, nations, continents, and the global society can all benefit. The findings of this study are recommended to climate stakeholders to help mitigate the effects of greenhouse gases on global warming.

Ethical Considerations: Ethics declaration not applicable.

Conflict of Interest: There is no conflict of interest.

Data Availability: Data obtained from the ‘carbon monitor’ web platform - carbonmonitor.org

REFERENCES

1. “Air Pollution.” World Health Organization, World Health Organization, (2019).www.who.int/health-topics/airpollution#tab=tab_1.
2. Babu Saheer L, Bhasy A, Maktabdar M and Zarrin J (2022). Data-Driven Framework for Understanding and Predicting Air Quality in Urban Areas. Front. Big Data 5:822573. doi: 10.3389/fdata.2022.822573
3. Bhardwaj J. and Sharma P. (2021). Artificial Intelligence-Based Smart Solution to Reduce Respiratory Problems Caused by Air Pollution. Journal of Emerging Investigators | vol 4.
4. Castelli M., Clemente F. M., Popovic A., Silva S., and Vanneschi L. (2020). A Machine Learning Approach to Predict Air Quality in California. Hindawi Complexity Volume 2020, Article ID 8049504, 23 pages https://doi.org/10.1155/2020/8049504.
5. Ghorani-Azam A, Riahi-Zanjani B, and Balali-Mood M. (2016) Effects of air pollution on human health and practical measures for prevention in Iran. J Res Med Sci.
6. Heydari A., Nezhad M. M., Garcia D. A., Keynia F., and De Santoli L. (2021). Air pollution forecasting application based on deep learning model and optimization algorithm. Springer.
7. Liu, Y., Jing Y., and Lu Y. (2020). Research on Quantitative Remote Sensing Monitoring Algorithm of Air Pollution Based on Artificial Intelligence. Hindawi, Journal of Chemistry, Volume 2020, Article ID 7390545, https://doi.org/10.1155/2020/7390545
8. Meivel, S., Mahesh, M., Mohnish, S., and Mukesh P, S. (2021). Air Pollution Monitoring System Using IOT And Artificial intelligence. International Journal of Modern Agriculture, Volume 10, No.2 ISSN: 2305-7246
9. Molinara, M., Ferdinandi, M., Cerro G., Ferrigno L., and Massera, E. (2019). An end-to-end indoor air monitoring system based on machine learning and SENSIPLUS platform. DOI 10.1109/ACCESS.2020.2987756, IEEE Access.

10. Sanjeev D. (2021). Implementation of Machine Learning Algorithms for Analysis and Prediction of Air Quality. International Journal of Engineering Research & Technology (IJERT). ISSN: 2278-0181 IJERTV10IS030323 Vol. 10 Issue 03.
11. Wood D., A. (2022). Local integrated air quality predictions from meteorology (2015 to 2020) with machine and deep learning assisted by data mining. Sustainability Analytics and Modeling 2 (2022) 10000.