Forecasting of Air Quality with Machine Learning

Daniel Samwel Makala[1,2]

[1]*Ministry of Energy Tanzania*
[2]*Department of Computer Science, China University of Petroleum (East), Qingdao, China*
Daniel.makala@outlook.com,

Abstract

In the age of increasing urbanization, industrial activities, and burgeoning population, the prediction of air quality has emerged as a critical concern. So far, different predictive models have been proposed for this purpose, from traditional statistical methods to Machine Learning models. This study investigates the advanced machine learning models, Support Vector Machine, and Long Short-Time Memory in the air quality prediction using hourly air quality index data from Dali, Taiwan. The LSTM shows its best performance compared to the SVM, achieving an RMSE of 3.7 and MAE of 2.4 compared to the SVM, which has RMSE and MAE of 43.24 and 32.4 respectively. These results show how the LSTM is powerful in capturing the dependencies in time series data. This study's findings underscore the potential of AI methods like LSTM as one of the powerful tools to be used in the prediction of time series like air quality. More to this, it opens doors for more research on machine learning applications in the environment and emphasizes that AI tools are the best-fit models. Overall, the above discussion has shown that LSTM models hold promise and potential for air quality forecasting.

Introduction

Air quality prediction has attracted much attention in environmental monitoring in recent decades. The direct reason for this is that the contents of air pollutants, especially the concentrations of fine particulate matter and special gases hidden in them, vary widely and pose a serious threat to public health [2]. In the age of increasing urbanization, industrial activities, and burgeoning population, the prediction of air quality has emerged as a critical concern. The rapid development of the economy and the sharp increase in urbanization have caused large amounts of industrial activities in various regions of the world. As a result, air pollution has become more severe throughout the world. Air pollutants can cause various damages and hazards to socioeconomics, ecosystems, and human health. Modal urban and city activities have greatly contributed to the deterioration of urban air quality [1]. Air pollution may bring about acute hazards to humans, such as an increase in the number of people suffering from asthma. It is estimated that more than 3 million premature deaths each year occur because of outdoor air pollution, mainly due to damage to the respiratory system.

Therefore, effective implementation of air quality prediction is an urgent priority. Fortunately, continuous technological advancements have remarkably benefited the fields of air quality monitoring and data collection. More advanced instruments and techniques are being applied to accurately and effectively acquire air pollutants over space and time. The implementation of air quality monitoring data support and public awareness over the past few years has resulted in the adjustment of environmental and health policies, which have successively declined adverse effects. The importance of air quality improvement has already been recognized. Significant efforts in model implementation and advanced data science models are required in research as well as policy formulation over a wide range. In the current exposition, it is feasible to predict air quality at a higher data resolution due to the collection of air quality information through scientific and effective environmental monitoring technologies. Predictive models can provide a

clear and objective forecast for effective policy implementation. However, developing effective air pollution prediction methods is challenging because of the complex non-linear, and high-dimensional systems and the dynamic changes in air quality components. Because of these challenges, the application of the Artificial intelligence model comes into place.

In this study paper, machine learning models of SVM and LSTM are used for the prediction of the air quality, since these models can handle the large volume of data as well as the non-linearity of the trend of air quality components.

Related Work

Air quality assessment has become increasingly important as the environment is increasingly deforested and industrialized with the population and expansion of cities. A variety of research has been conducted to predict the future quality of the air in different cities and nations across the globe [3]. However, due to the trend and characteristics of the air quality index data, artificial intelligence models have been doing very well. AI models (machine learning and deep learning) have been doing very well in the prediction of different volatile data such as oil prices, gold, and others. [1], [2], [3], [4].

When conducting studies on the prediction of air quality in the Indian coastal city of Visakhapatnam using machine learning techniques, Ravindran and his colleagues applied five models for better results. They employed LightGBM, Random Forest, Cat Boost, Adaboost, and XGBoost. Their practical results show that the model Cat Boost model outperformed other models by having an r-square correlation coefficient of 0.9998, a Mean Absolute Error (MAE) of 0.60, a Mean Square Error (MSE) of 0.58, and a Root Mean Square Error (RMSE) of 0.76. The Ad boost model had the least effective prediction with an R square correlation coefficient of 0.9753 [5]. In addition to that [6] Gupta et.al proposed and analyzed three methods. SVR, Random Forest regression, and Cat Boost regression in the prediction of Air Quality in New Delhi, Bangalore, Kolkata, and Hyderabad. Their result indicates that Random Forest has performed better in Bangalore, Kolkata, and Hyderabad.

There are numerous the studies and researches that have been conducted using different models traditional and machine learning models. In most cases, the machine learning model has been showing the most promising performance, by having a high prediction accuracy rate. [7], [8], [9].

Methodology

Data Collection

To predict changes in air quality, the first and foremost step is the availability of sufficient data on the ground. There are various sources from which large-scale data for air quality can be obtained, including satellite measurements, ground stations, surface and underground transportation systems, weather forecasts, industrial sources, and sensor measurements. The data collection preprocessing step exhibits noticeable stress in the predictive analysis of the air quality. The data used in this study has been collected from various sources. [10] [11].

Data Processing and Cleaning

Data preprocessing includes cleaning, normalization, and dimensionality reduction procedures that aim to obtain the cleaned input data. Although data obtained from a single source might provide incomplete or biased information, a combination of sources can present the data featuring high relevance and input quality. Cleaning the raw data is meant to root out noisy and irrelevant information collected from

different sources, ensuring that the input data is relevant and does not contain conflicting or missing information. Unclean data may contain invalid values or missing data that can cause wrong predictions.

SVM Architecture

Support Vector Machines (SVM) have been widely applied as an effective tool in many predictive tasks and decision-making problems. SVM constructs a hyperplane or a set of hyperplanes in a high infinite-dimensional space, which can be used for both classification and regression. The basic idea of SVM is to find the optimal hyperplane that separates a set of instances in the input space by considering the hyperplane with the maximal margin. In SVM it uses a function known as a kernel function to transform the data into a high-dimensional space, where a linear hyperplane can be used to separate the classes. These kernels are linear kernel, polynomial kernel, radial basis function, and sigmoid kernel. The idea of SVM can be used in engineering applications as well as complex problems in environmental science and modeling. Unlike the feedforward neural networks that adopt a black-box approach to map input-output relationships, the decision rule of an SVM model can be interpreted and explained based on the structural risk minimization principle. Graphically SVM can be presented as in **Figure 1**.

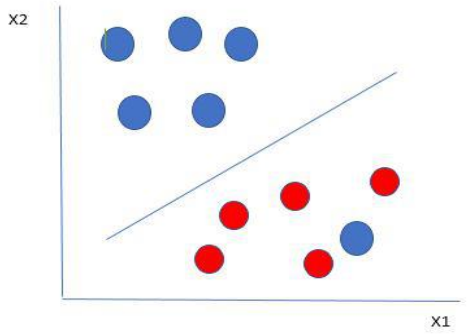


Figure 1 Showing best-fit hyperplane [12].

LSTM Architecture

LSTMs are a special kind of Recurrent Neural Networks (RNNs) that are well-suited for sequential data, as they can model temporal dependencies and patterns. [13]LSTMs can capture the complexities and influences in the sequence and the variations in the values. Furthermore, they are capable of learning the relationship of patterns within time series data regardless of the gap between them. In time series data, especially in air quality data, it becomes hard to find and fix the pattern due to the high-level interference of human activities.

LSTM is an updated RNN with a memory cell that maintains all the information passed through. Additionally, it consists of three gates where all the functionality is done through these gates. These gates are:

1. **Input Gate**: Here is where the decision of which information to add to the cell state. It also has a sigmoid function that controls which value is to enter the state of LSTM. The output of the input gate is multiplied with tanh, producing a new value

2. **Forget Gate:** here is where the decision on what information to discard is made. The information from the previous state and the current input pass through the function known as sigmoid activation and the output value is in the range of 0 to 1, where the values close to 1 are retained and closer to o are discarded.

3. Output gate: here is where the decision to output the information takes place. The sigmoid function is applied to determine the necessity of the information to be output or to continue to another cell state.

Generally, the architecture of LSTMs includes multiple memory cells that facilitate maintaining information across different time steps. These memory cells rest within each unit and take the form of internal self-connected units with a linear activation function. The Structure of LSTM can be seen in Figure 2.

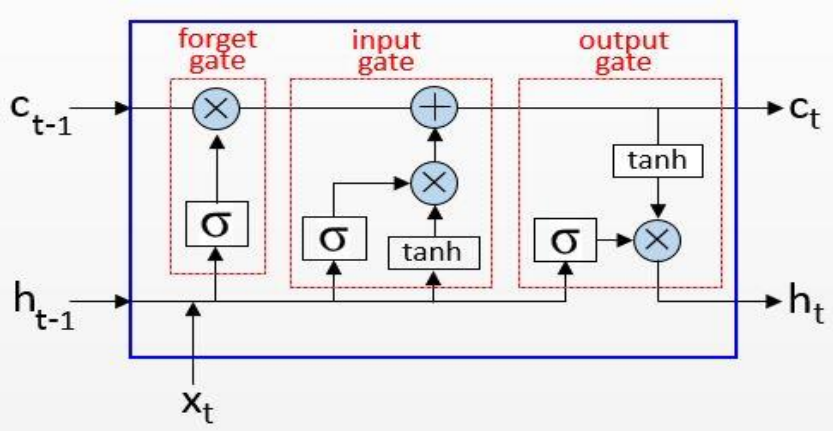


Figure 2 General architecture of LSTM block cell [14]

Performance Metrics

It is necessary to evaluate the performance of how the models have been performed. There are different ways of measuring how the models have performed. In this study, three metrics have been used. The first metric is Root Mean Square Error. RMSE is the square root of the average squared difference between actual and prediction values.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(predict - Actual)^2}{N}}$$

The second method is Means Absolute Error (MAE). It is the absolute average distance between the Predicted and original values. Un RMSE and MAE, the lower the value, the better the performance.

$$MSE = \sqrt{\frac{\sum_{i=1}^{n}(predict - Actual)^2}{N}}$$

Lastly, there is R-squared, which explains how the predicted value closely matches the true value. The R-squared value is between 0 and 1, where 1 is 100% matches.

$$R - Squared = = 1 - \left(\frac{SS_{reg}}{SS_{Total}}\right)$$

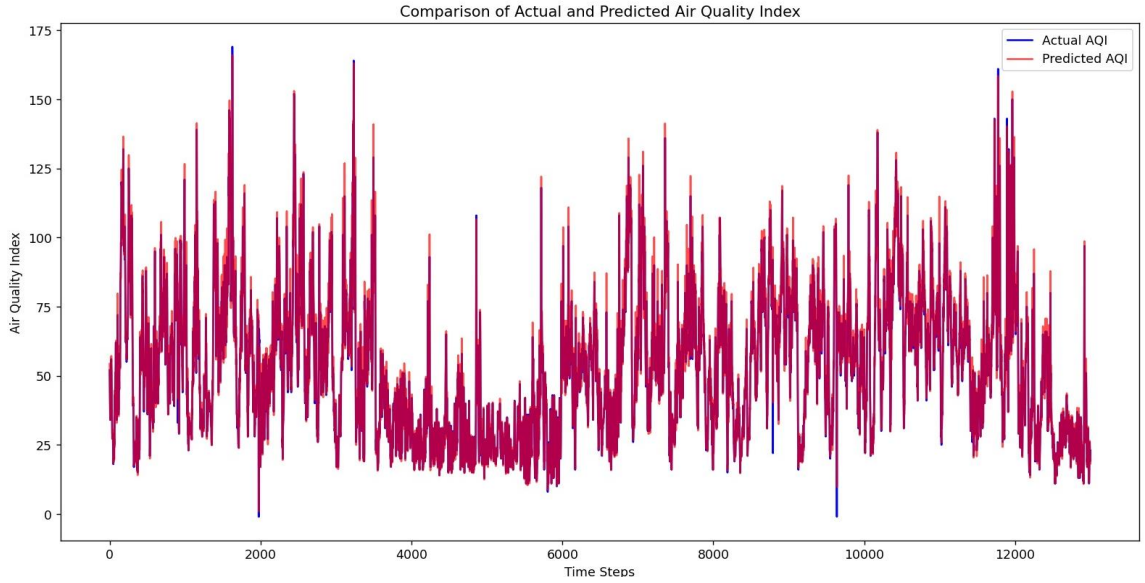
Results and Discussion

The application of deep learning makes a pioneering contribution to air quality prediction by enhancing early warning mechanisms. For this research study, the data are split into the training data and testing data. The training data comprises data from January 2017 to December 2022. The rest are for testing the model.

Table 1

	RMSE	**MAE**	**R-Squared**
LSTM	3.7	2.4	98%
SVM	43.24	32.4	19%

The results of this study demonstrate the remarkable performance of the LSTM in handling the time series data. As shown in Table 1, the LSTM achieved an RMSE of 3.7 and MAE of 2.4, which implies the prediction results are very closely aligned in LSTM compared to the SVM where it has RMSE and MAE of 43.24 and 32.4 respectively.



Additionally, the R square of the LSTM model is 98%, which implies a better level of accuracy in the variance of the data. When the value of the R-square is close to 1, (100%) it shows the prediction values match the true values of the air quality. This also can be seen in Figure 3, which shows the graph of predicted values against the true values in LSTM.

Conclusion

The advancement of machine learning and deep learning technologies has the potential to empower not only environmental scientists and professionals but the wider public by characterizing pollution dispersion paths and informing precautionary health advice. Accurate forecasting of the air quality forecast is a pivotal element of public health management programs in modern, rapidly growing metropolises. The continuous growth rates of urban environments, the increase in vehicle usage, and industrial activity have a direct unfavorable impact on air quality.

Reference

[1] D. Makala and Z. Li, "ECONOMIC FORECASTING WITH DEEP LEARNING: CRUDE OIL," *MATTER: International Journal of Science and Technology*, vol. 5, no. 2, pp. 213–228, Oct. 2019, doi: 10.20319/mijst.2019.52.213228.

[2] A. Gadhave, "Gold Price Prediction using Machine Learning," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 06, no. 05, Jun. 2022, doi: 10.55041/IJSREM15027.

[3] G. S. Vidya and V. S. Hari, "Gold Price Prediction and Modelling using Deep Learning Techniques," *2020 IEEE Recent Advances in Intelligent Computational Systems, RAICS 2020*, pp. 28–31, Dec. 2020, doi: 10.1109/RAICS51191.2020.9332471.

[4] Z. Yu, K. Wang, Z. Wan, S. Xie, and Z. Lv, "Popular deep learning algorithms for disease prediction: a review," *Cluster Comput*, vol. 26, no. 2, pp. 1231–1251, Apr. 2023, doi: 10.1007/S10586-022-03707-Y/FIGURES/3.

[5] G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, and C. Sonne, "Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam," *Chemosphere*, vol. 338, p. 139518, Oct. 2023, doi: 10.1016/J.CHEMOSPHERE.2023.139518.

[6] N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis," *J Environ Public Health*, vol. 2023, no. 1, p. 4916267, Jan. 2023, doi: 10.1155/2023/4916267.

[7] Y. Rybarczyk, R. Zalakeviciute, Y. Rybarczyk, and R. Zalakeviciute, "Regression Models to Predict Air Pollution from Affordable Data Collections," *Machine Learning - Advanced Techniques and Emerging Applications*, Dec. 2017, doi: 10.5772/INTECHOPEN.71848.

[8] G. I. Drewil and R. J. Al-Bahadili, "Air pollution prediction using LSTM deep learning and metaheuristics algorithms," *Measurement: Sensors*, vol. 24, p. 100546, Dec. 2022, doi: 10.1016/J.MEASEN.2022.100546.

[9] B. S. Freeman, G. Taylor, B. Gharabaghi, and J. Thé, "Forecasting air quality time series using deep learning," *J Air Waste Manage Assoc*, vol. 68, no. 8, pp. 866–886, Aug. 2018, doi: 10.1080/10962247.2018.1459956.

[10] "Dataset - Catalog." Accessed: Jan. 01, 2025. [Online]. Available: https://catalog.data.gov/dataset/?tags=air-quality

[11] “Find Open Datasets and Machine Learning Projects | Kaggle.” Accessed: Apr. 10, 2024. [Online]. Available: https://www.kaggle.com/datasets?search=gold+price

[12] “Support Vector Machine (SVM) Algorithm - GeeksforGeeks.” Accessed: Jan. 02, 2025. [Online]. Available: https://www.geeksforgeeks.org/support-vector-machine-algorithm/

[13] L. L. S. Sachetti and V. F. S. Mota, “Time Series Prediction about Air Quality using LSTM-Based Models: A Systematic Mapping”, Accessed: Jan. 02, 2025. [Online]. Available: https://ieeexplore.ieee.org

[14] “function of out gate in lstm - Google Search.” Accessed: Jan. 02, 2025. [Online]. Available: https://www.google.com/imgres?q=function%20of%20out%20gate%20in%20lstm&imgurl=https%3A%2F%2Fdwbi1.wordpress.com%2Fwp-content%2Fuploads%2F2021%2F08%2Ffig-4-lstm.jpg&imgrefurl=https%3A%2F%2Fdwbi1.wordpress.com%2F2021%2F08%2F07%2Frecurrent-neural-network-rnn-and-lstm%2F&docid=T4q91EZnCi2_FM&tbnid=ATw-9WeQ5-y-tM&vet=12ahUKEwjchZyN79aKAxUUTGwGHRiRBrwQM3oECGkQAA..i&w=486&h=281&hcb=2&ved=2ahUKEwjchZyN79aKAxUUTGwGHRiRBrwQM3oECGkQAA