Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Multi-Scale Temporal Convolutional Networks for Long-Range PM2.5 prediction in Taiwan's Monsoon-Influenced Climate

Vivek Bongale[1], HarishKumar K S[2]
12Big Data Laboratory, School of Computer Science Engineering, Presidency University Bangalore Vivek.bongale@presidencyuniversity.in, Harishkumar@presidencyuniversity.in

ARTICLE INFO

Received: 24 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

One of the most important components of the environment is air. The earth's temperature, ecosystems, human health, and environmental sustainability are all at risk due to the growing global air pollution catastrophe. Because of how sneaky it is, air pollution has been called a silent killer. Its harmful effects are further highlighted by its indirect impacts on human health. Millions of lives could be saved worldwide if air quality is detected early. Consequently, there is a lot of interest among researchers in the analysis and prediction of air pollution. Neural networks, deep learning, and conventional machine learning are among the research topics. The problem of correctly and efficiently forecasting air pollution becomes crucial. Using ML, DL, and neural network techniques, research aims to forecast and make prediction on PM2.5 concentration at the MCMUG station in Taiwan. Based on the LSTM deep learning method, Random Forest ML algorithm, ANN algorithm, and Gradient Boosted Model which incorporates the XGBoost, LightGBM, and CATBoost techniques? As inputs to the model, the Taiwan Air Quality Monitoring Board's traffic statistics, pollutant information, and climatic characteristics from January 1, 2018, to December 31, 2023 were examined. These are the models beat the predictive performance of existing conventional models. Statistical indicators like RMSE, MAE, and MSE respectively. R2 the co-efficient determinant will be used to evaluate the MCMUG station the effectiveness of various strategies.

Keywords: ML and DL Prediction model, Random forest, PM2.5, Air Quality Index, Air Pollutants

INTRODUCTION

Air pollution is becoming more of a concern as a result of its major effects on both human health and climate change. The environment, human existence, job productivity, and energy efficiency are all negatively impacted by the chemical emissions from human activities, such as the use of aerosol products and the burning of fossil fuels. To effectively regulate air contaminants, regular monitoring is necessary [1]. The usage of energy and human activity cause air pollution, which releases hazardous chemicals into the atmosphere through sources such as burning coal, kerosene, straw, and other fossil fuels. These pollutants have negative impacts on plant development, human and animal health. One of the main causes of air pollution is the AQI, a linear measure of the concentration of pollutants. Recent improvements in technology and advanced sensors have greatly enhanced the accuracy and efficiency of air quality monitoring, notwithstanding past difficulties. This is essential for efficient regulation and for tackling the complicated link between climate change and the economic viability of agriculture.

The main cause of global warming is greenhouse gases, which have a detrimental effect on the economy, the environment, and agriculture. According to the World Health Organization's 2022 study on air quality, PM2.5 levels are rising globally. Taiwan has some of the worst air pollution in the world, with 20 of its 25 most contaminated cities. The intricate connections between air quality and AQI categories have been deciphered using machine learning models. Ground-based sensor networks and satellite observations, two common methods, have drawbacks, but sophisticated statistical modeling approaches provide a more complete picture. The Pollution Standards Index (PSI) in Taiwan assesses air quality in locations that are susceptible to severe pollution. Modern

Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

ML and DL models, such as the Random Forest Regression Algorithm, ANNs model, LSTM, and Gradient Boosted models that include the XGBoost, LightGBM, and CATBoost techniques [2], have been used to better understand pollution variability, identify links between meteorological variables, and predict future pollution levels.

The levels of particulate-matter and several contaminants, such as NOx, CO, PM2.5, O3, SO2, PM10 etc., are highly correlated. This gives the air pollution level as well as the air pollution particle matter level. PM2.5 particles are minute than 2.5µg and have been connected with a variety of harmful human health's impacts, like cardiovascular and respiration related disease. Consequently, health depends on PM2.5. It is now obvious and imperative to plan for and create an early warning system that will give citizens air quality updates. The major goals and motivation of smart city is to respond to sensor data. However, sensors may sometimes fail and give erroneous data. In smart cities to forecast the air quality, predictive models seems to be a potential remedy to these problems. The main goal to compare various ML techniques used to forecast PM2.5. Consequently, this model can predict when sensors will fail with the least amount of error in the level of PM2.5 in the air, allowing it to issue an aware when particular threshold values are reached.

The suggested ML algorithms are used in this study to forecast PM2.5 concentrations. We collected data from the TAQMN in Taiwan for the MCMUG station between January 1, 2018, and December 31, 2023, to train the model. The TAQMN provides both meteorological and air pollution data.

LITERATURE SURVEY

Air pollution is a major issue worldwide that has a major adverse effect on the environment and on human health's. Traditional fixed stations, which are now used for air quality monitoring, have inherent drawbacks like high expenses, limited geographical coverage, and a lack of real-time data availability. However, cutting-edge technologies, particularly ML and the IoT, have opened the door for innovative solutions that improve forecast accuracy and enable live monitoring. The efficiency and quality of air control actions is increased by using IoT-based systems, which offer a dynamic approach that provides real-time data with improved spatiotemporal resolution [3]. The prediction of the AQI has made notable progress thanks to the implementation of complex ML and DL models. Effectiveness of these algorithms in the field of air quality forecasting has been highlighted by the remarkable precision of algorithms like random forest method and SVM in predicting AQI levels. Additionally, convolutional-neural-networks (CNNs) and LSTM methods are two examples of DL techniques that are significantly apt at predicting time series, which allows for a more sophisticated analysis of changes in air quality over time. By employing ensemble learning approaches, such as those seen in models like Random-Forest and XGBoost, we have demonstrated improved predictive abilities over single models, paving the way for more accurate air quality predictions. In particular, monitoring data and secondary modeling methodologies have been used with unmatched accuracy by cutting-edge machine learning algorithms like CatBoost and LightGBM, which have become leaders in air quality class prediction [4]. Ultimately, the integration of cutting-edge technologies into air quality monitoring represents a major step forward in lessening the negative consequences of air pollution on public health and environmental sustainability.

DATASET DESCRIPTION

The research focuses on the Air Quality Index (AQI) in Mcmug (Magong) city, located in the Penghu Islands, Taiwan. The city sits at latitude 23° 33' 55. 44" N and longitude 119° 35' 10. 57" E. Data was collected from the TAQMN between 2018 and 2023, covering 76 stations across different locations. Magong is the largest city in Penghu [5], with an area of 33. 99 km², and serves as a center for industries, tourism, fishing, and light manufacturing. It is one of the fastest-growing cities in Taiwan and a popular tourist destination. As shown in Figure 1:

Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

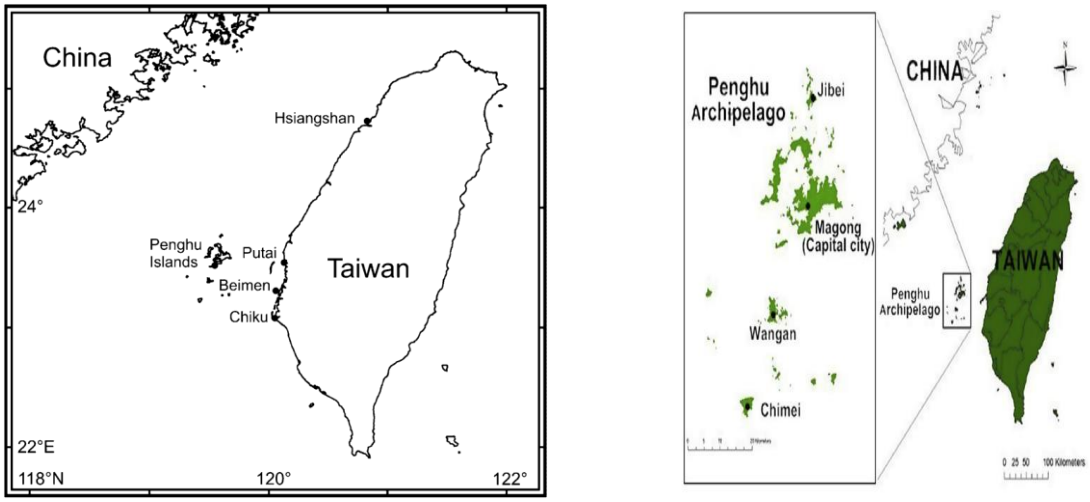

Figure 1: Location map of the MAGONG, PENGHU Island of TAIWAN

DATASET PRE-PROCESSING

From January 2018 to December 2023, a total of 2191 PM2. 5 instances were collected from open source data by TAQMN. Any missing values in the dataset were removed, resulting in these 2191 instances. The dataset for AQI underwent a type conversion from object to float data. Performance evaluation of different models was done by dividing the dataset into 80% for train dataset and 20% for test dataset [6]. An auto-correlation noise engineering technique was applied to ensure the index was sorted before using a rolling window to address any remaining missing values. This also handled cases where the first six smoothed values might be NaN. The data was normalized using MinMaxScaler. The models' forecasting ability for AQI was assessed with metrics like RMSE, MSE, MAE, and R^2.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2192 entries, 0 to 2191
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Unnamed: 0    2192 non-null   int64
 1   date          2192 non-null   object
 2   Station       2192 non-null   object
 3   PM2.5         2192 non-null   float64
dtypes: float64(1), int64(1), object(2)
memory usage: 68.6+ KB
(   Unnamed: 0        date Station      PM2.5
 0           0  01-01-2018   McMug  41.291667
 1           1  02-01-2018   McMug  63.875000
 2           2  03-01-2018   McMug  44.916667
 3           3  04-01-2018   McMug  29.541667
 4           4  05-01-2018   McMug  20.500000,
 None)
```

Figure 2: Sample Dataset and Applied Feature Engineering on MCMUG Station

METHODOLOGY AND RESULT ANALYSIS

The Air Quality Index of the Mcmug city was predicted using ML algorithms like Random Forest, Artificial Neural Networks (ANNs) model, LSTM, and gradient boosted models that integrate the XGBoost, LightGBM, and CATBoost methodologies [7]. The datasets used to calculate the AQI's organization as shown in Figure2.

Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

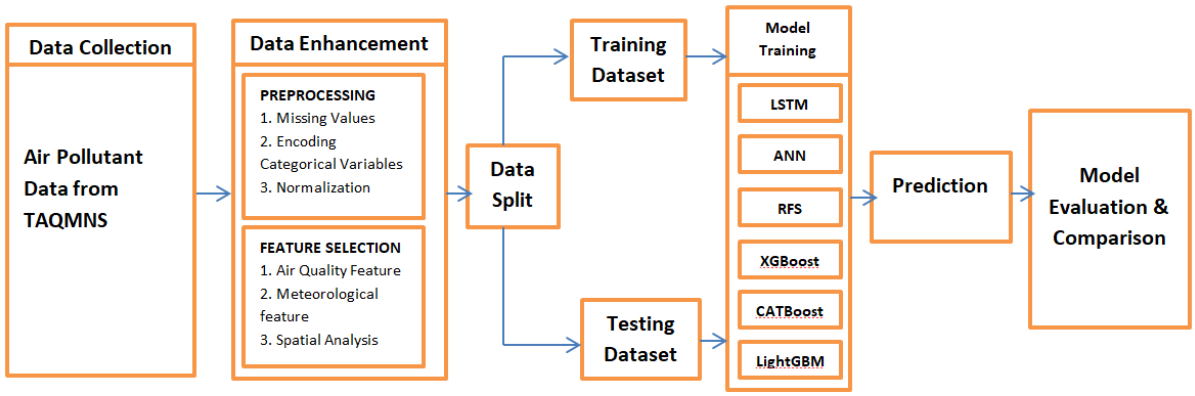

Figure 3: Architecture for AQI prediction using ML algorithms

LSTM:

LSTM, a type of RNNs algorithm, concentrates on the vanishing gradient issues in conventional RNNs, making it ideal for PM2.5 forecasting as it is able to identify long-term dependencies.

It basically feed sequences into the LSTM, which learns to map input sequences to the next day's PM2.5 value [8]. To train the model it uses a MSE i.e., a loss function and an optimizer. Compute the performance metrics (R^2, MSE, RMSE, MAE, AE) on train dataset and test datasets to assess models accuracy.

In order to predict the next day's PM2.5 value, hidden state ht is typically passes through a fully connected layer:

$$y_t = W_y \cdot h_t + b_y \quad ------(1)$$

Where:

y_t : Predicted $PM_{2.5}$ value (normalized)

W_y, b_y : Weight matrix and bias for the output layer.

$PM_{2.5}$: y_t is inverse-transformed (e.g., using MinMaxScaler) to obtain the actual $PM_{2.5}$ concentration in $\mu g/m^3$.

ANN (Artificial Neural Network):

The anatomy and functioning of biological neurons serve as the inspiration for a computational model known as an ANN algorithm. Through a process of weighted connections, activation functions, and optimization, it is made up of interconnection of nodes which is neurons and it is arranged in layers with the goal of learning patterns from data [9]. Because ANNs can model non-linear interactions, they are frequently employed for regression tasks like forecasting air pollution concentrations like PM2.5.

The core equation for a single neuron in an Artificial Neural Network (ANN) used for prediction is:

$$Z = W1X1 + W2X2 + ...WnXn + b \quad ------(2)$$

Where:

Z: represents the output of the neuron.

W1, W2, ..., Wn: are assigned weights to each input's (X1, X2, ..., Xn).

X1, X2, ..., Xn: are input's values.

b: is bias term (also known as the intercept).

Random-forest Model:

Random Forest is an ML technique that enhances prediction reliability and precision by combining multiple decision trees, reducing over fitting and providing consistent predictions through averaging.

Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Random Forest for regression predicts a continuous output by considering the averages of multiple decision trees predictions [10]. Below are the equations for prediction:

$$y = \frac{1}{T}\sum_{t=1}^{T} ht(x) - - - - - - - - - - - (3)$$

Where:

y = Predicted PM2.5 value (normalized, [0, 1]).

T = No. of trees

ht(x) = t^{th} decision tree for input x prediction

XGBoost:

XGBoost is a machine learning technique that uses weak learners to build prediction models, making it ideal for regression applications like PM2.5 concentration prediction, and regularization to prevent over fitting [11].

The learning rate determines the scaling of the sum of predictions from all trees in the XGBoost prediction:

$$y = \sum_{t=1}^{T} n.ft(x) - - - - - - - - - - - - - (4)$$

Where:

ft(x) : Output of the tth decision tree, which maps input x to a leaf node with a weight wj,t .

n: Learning rate (e.g., 0.1), controlling the contribution of each tree to prevent overfitting.

y : Predicted PM2.5 value (normalized, [0, 1]).

CATBoost Model:

CatBoost is a gradient boosting method designed for handling categorical information, enhancing performance in regression and classification tasks [12]. It incorporates ordered boosting and native categorical factor handling, making it suitable for time-series data with structured inputs.

LightGBM Model:

LightGBM is a gradient boosting framework that scales well and is designed for maximum efficiency in terms of computation and memory utilization [13], making it especially useful for predicting PM2.5 levels. It employs histogram learning based and leaf-wise tree growth for increased accuracy and speedier training.

Performance Criteria:

Performance of the models is assessed by utilising the number of statistical measures, which includes R^2, MAE, MSE, and RMSE [14]. The following displays the criteria formulas:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} - - - - - - - - - - - - (5)$$

Where, y_i shows the actual $PM_{2.5}$ observed value, $\hat{y}_i$ Predicted $PM_{2.5}$ value and $\bar{y}_i$ Mean of Actual $PM_{2.5}$ values [15].

Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad ----------(6)$$

Where, n shows the no. of observations $\hat{y}_i$ is the predicted value, and y_i says the actual value.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad ----------(7)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad ----------(8)$$

RESULTS AND DISCUSSIONS:

This study took into account the stationary time series data from the TAQMN dataset spanning Jan 2018 to Dec 2023, which also considered the main cities of Taiwan: Mcmug, annan, Good, chiayi and Newport. The Mcmug station that is chosen and exclusively concentrates on PM2.5 since it is the major detrimental consequence of airborne pollution. The information was gathered hourly and subsequently transformed into daily and monthly formats. Similarly, several models, such as the ANN, LSTM, RFS, XGBoost, CATBoost, and LightGBM, were used to test the suggested model. Different models are used to conduct the tests, and for each algorithm, the cross validation and performance metrics are utilised to assess its precision. Connection between real and expected values is illustrated by the determination coefficient.

Table1: The optimal outcome of the diagnostics training and testing dataset values for various ML algorithm models to predict the PM2.5

	Training Dataset				**Testing Dataset**			
Methods	**R^2**	**MSE**	**RMSE**	**MAE**	**R^2**	**MSE**	**RMSE**	**MAE**
LSTM	0.9271	2.2407	1.4969	0.9754	0.7719	0.8923	0.9446	0.6960
ANN	0.9213	2.3433	1.5308	1.0843	0.6882	1.2488	1.1175	0.8855
RFS	0.9640	0.150	0.387	0.290	0.8970	0.2440	0.4940	0.4030
XGBoost	**0.9990**	**0.0296**	**0.1722**	**0.1338**	**0.9851**	**0.0475**	**0.2180**	**0.1691**
CATBoost	0.8747	7.7290	2.7801	2.1643	0.3253	9.0552	3.0091	2.3422
LightGBM	0.8691	7.3588	2.7127	1.9727	0.3664	6.7023	2.5888	1.9523

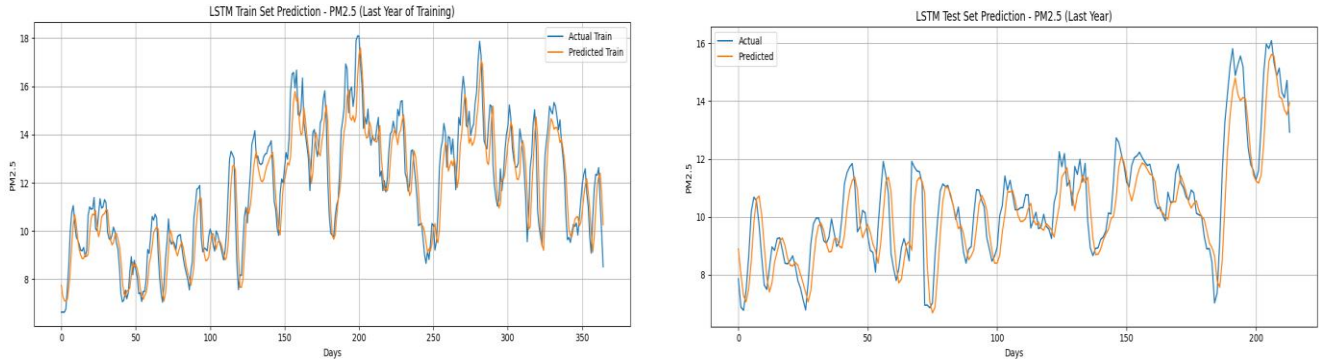

Figure 4: Prediction for **LSTM** Training and Testing Dataset for Mcmug Station.

Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

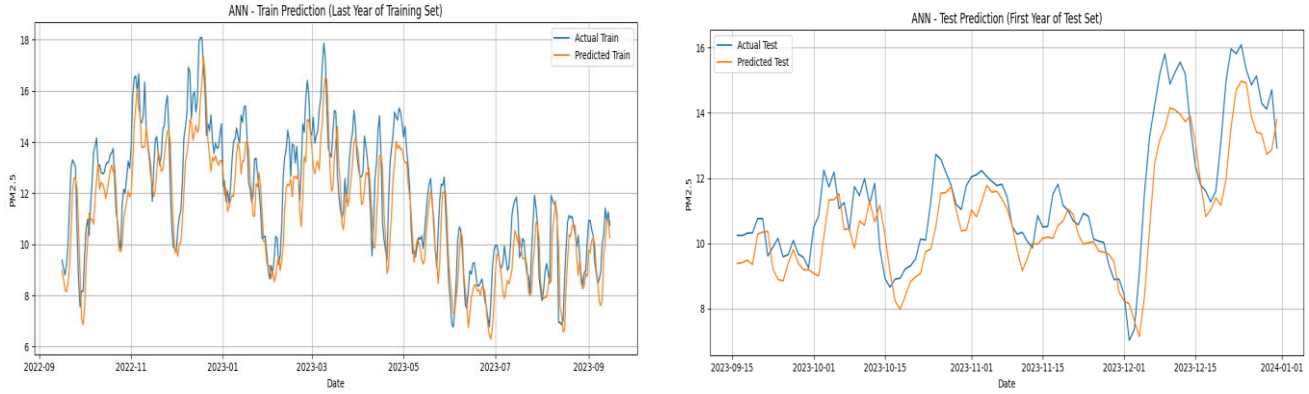

Figure 5: Prediction for **ANN** Training and Testing Dataset for Mcmug Station.

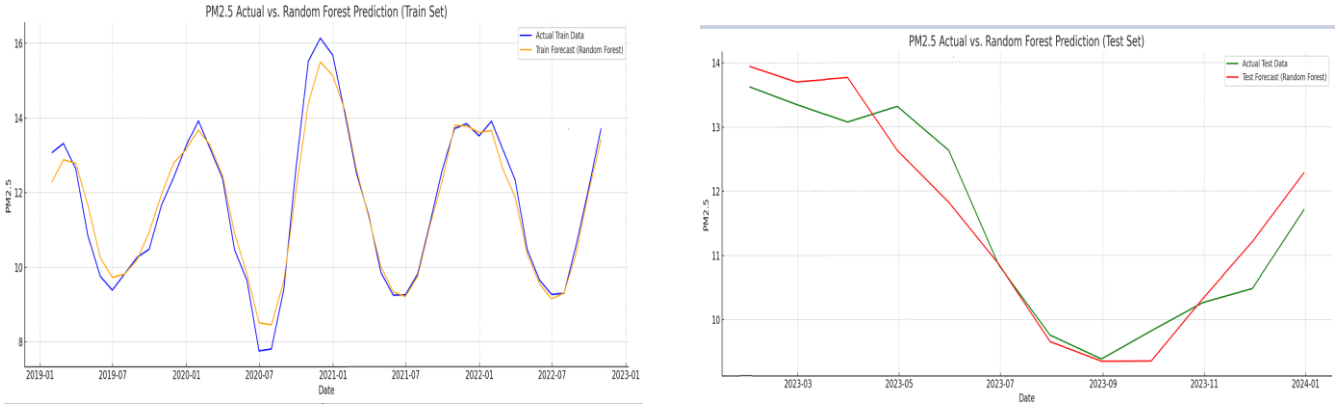

Figure 6: Prediction for **RFS** Training and Testing Dataset for Mcmug Station

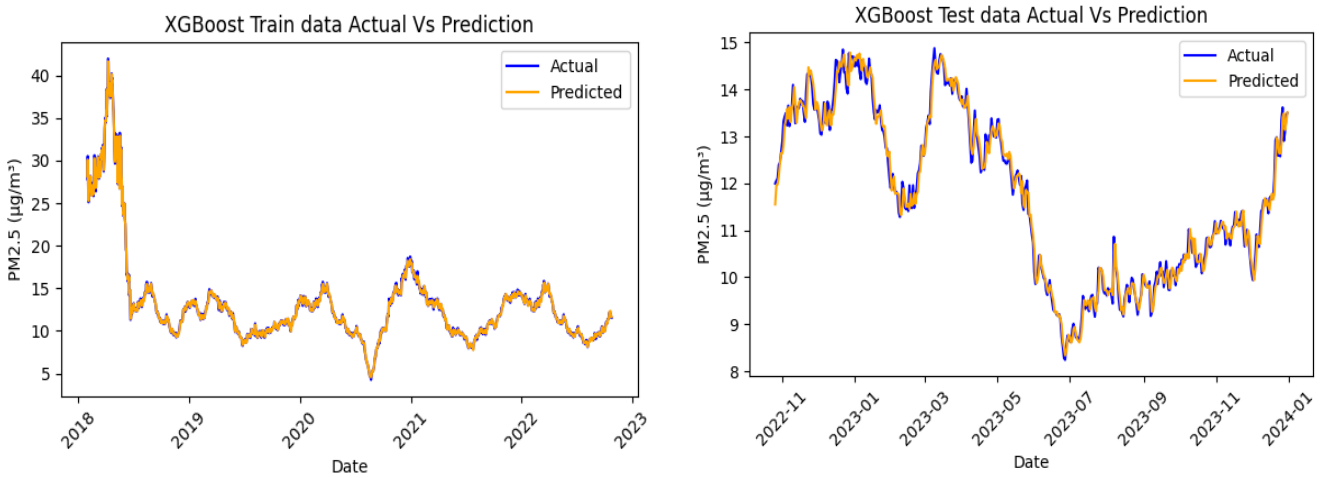

Figure 7: Prediction for **XGBoost** Training and Testing Dataset for Mcmug Station

Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

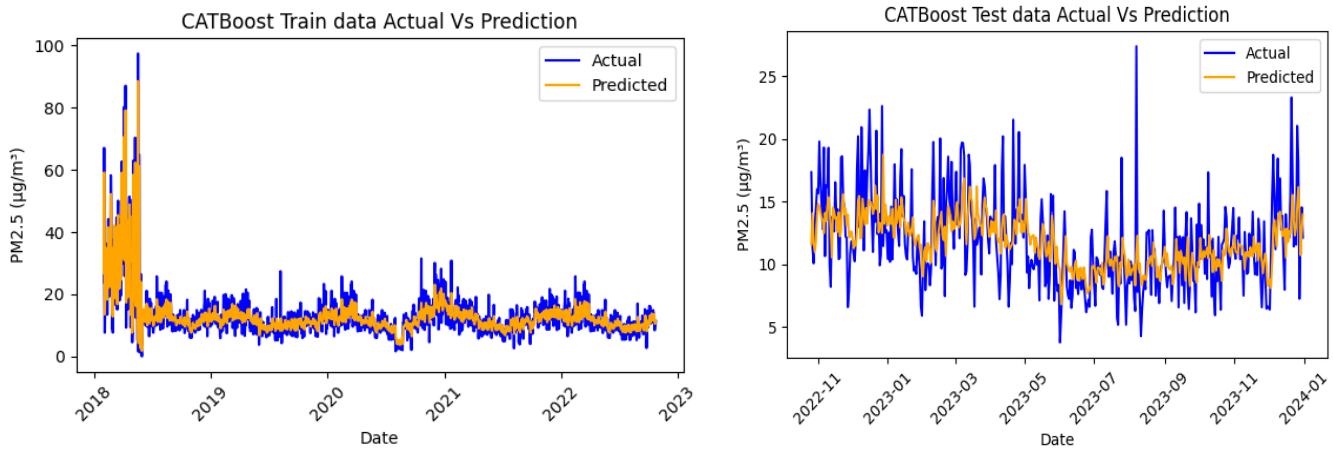

Figure 8: Prediction for **CATBoost** Training and Testing Dataset for Mcmug Station

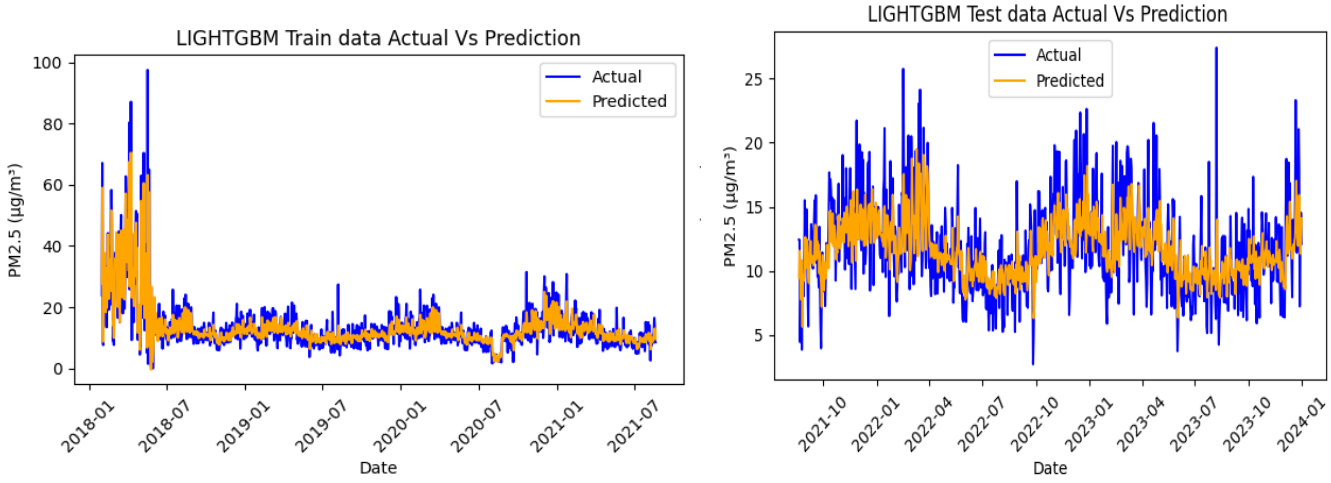

Figure 9: Prediction for **LIGHTGBM** Training and Testing Dataset for Mcmug Station

The Long Short Term Model (LSTM) prediction results are displayed in Figure 4; the train dataset is shown on the left, and the test dataset is shown on the right, with R^2 values of 0.9271 and 0.7719, respectively. Additionally, the Orange colour line signifies the particulate matter PM2.5 prediction, while the Blue colour line represents the actual values. In similar vein, Figure 5 displays the prediction outcomes for ANN algorithms; the R^2 values for train and test data are 0.9213 and 0.6882, respectively. The Random Forest algorithm's R^2 values for train and test data are 0.9640 and 0.8970, respectively, as shown in Figure 6. The XGBoost regression technique produced the best model for PM2.5 prediction when compared to the performance with the rest of other algorithms, as seen in Figure 7 and the coefficients of determination data are displayed in Table1, where the R^2 values of the train and test data were 0.9990 and 0.9851, accordingly. Figure 8 and 9 projects the R^2 value for the model CATBoost and LightGBM. Figure 10 shows the comparisons of the performance valuation matrix for different AI models on training and testing MCMUG dataset.

Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Figure 10: Comparisons of the Performance Valuation Matrix for different AI Models

CONCLUSION AND FUTURE WORK

The suggested machine learning models for analysing Taiwan's air pollution using TAQMN data are introduced in this paper. The TAQMN data from 2018 Jan to 2023 December specifically shows 76 air pollution stations. ML models based on statistical computations of metric's like MAE, MSE, RMSE, and R^2 are used to forecast particulate matter PM2.5. The findings demonstrate that the suggested models outperform the earlier models in terms of performance and that the projected and actual values are quite similar to one another. Ultimately, we determine that the XGBoost model performs better in predicting air pollution at Taiwan's MCMUG station. Future research should explore the specific sources of pollution at each station and assess the effectiveness of mitigation strategies to addresses the disparities and improves public health outcomes across Taiwan.

REFERENCES

[1] Chang-Hoi, Ho, Ingyu Park, Hye-Ryun Oh, Hyeon-Ju Gim, Sun-Kyong Hur, Jinwon Kim, and Dae-Ryun Choi. "Development of a PM2. 5 prediction model using a recurrent neural network algorithm for the Seoul metropolitan area, Republic of Korea." Atmospheric Environment 245 (2021): 118021.

[2] Alimissis, Anastasios, Kostas Philippopoulos, C. G. Tzanis, and Despina Deligiorgi. "Spatial estimation of urban air pollution with the use of artificial neural network models." Atmospheric environment 191 (2018): 205-213.

[3] Dotse, Sam-Quarcoo, Mohammad Iskandar Petra, Lalit Dagar, and Liyanage C. De Silva. "Application of computational intelligence techniques to forecast daily PM10 exceedances in Brunei Darussalam." Atmospheric Pollution Research 9, no. 2 (2018): 358-368.

[4] Harishkumar, K. S., and K. M. Yogesh. "Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models." Procedia Computer Science 171 (2020): 2057-2066.

[5] Imam, Mohsin, Sufiyan Adam, Soumyabrata Dev, and Nashreen Nesa. "Air quality monitoring using statistical learning models for sustainable environment." Intelligent Systems with Applications 22 (2024): 200333

[6] Devasekhar, V., and P. Natarajan. "Prediction of air quality and pollution using statistical methods and machine learning techniques." International Journal of Advanced Computer Science and Applications 14, no. 4 (2023).

[7] Morapedi, Tshepang Duncan, and Ibidun Christiana Obagbuwa. "Air pollution particulate matter (PM2. 5) prediction in South African cities using machine learning techniques." Frontiers in Artificial Intelligence 6 (2023): 1230087

[8] Ravindiran, Gokulan, Gasim Hayder, Karthick Kanagarathinam, Avinash Alagumalai, and Christian Sonne. "Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam." Chemosphere 338 (2023): 139518

[9] Rahman, Md Mahbubur, Md Emran Hussain Nayeem, Md Shorup Ahmed, Khadiza Akther Tanha, Md Shahriar Alam Sakib, Khandaker Mohammad Mohi Uddin, and Hafiz Md Hasan Babu. "AirNet: predictive

machine learning model for air quality forecasting using web interface." Environmental Systems Research 13, no. 1 (2024): 44.
[10] Liu, Qian, Bingyan Cui, and Zhen Liu. "Air quality class prediction using machine learning methods based on monitoring data and secondary modeling." Atmosphere 15, no. 5 (2024): 553.
[11] Alkhodaidi, Amjad, Afraa Attiah, Alaa Mhawish, and Abeer Hakeem. "The Role of Machine Learning in Enhancing Particulate Matter Estimation: A Systematic Literature Review." Technologies 12, no. 10 (2024): 198.
[12] Wu, Yong, Xiaochu Wang, Meizhen Wang, Xuejun Liu, and Sifeng Zhu. "Time-Series Forecasting of PM2. 5 and PM10 Concentrations Based on the Integration of Surveillance Images." Sensors 25, no. 1 (2024): 95.
[13] Zhang, Yuyi, Qiushi Sun, Jing Liu, and Ovanes Petrosian. "Long-term forecasting of air pollution particulate matter (PM2. 5) and analysis of influencing factors." Sustainability 16, no. 1 (2023): 19.
[14] Warren, Joshua L., Wenjing Kong, Thomas J. Luben, and Howard H. Chang. "Critical window variable selection: estimating the impact of air pollution on very preterm birth." Biostatistics 21, no. 4 (2020): 790-806.
[15] Henning, Robert J. "Particulate matter air pollution is a significant risk factor for cardiovascular disease." Current Problems in Cardiology 49, no. 1 (2024): 102094.