PM10-Based Air Quality Prediction Using Machine Learning for Environmental Monitoring

Gudipati Sri Vibhavan[1], Rajasekar J[2], Ande Umesh Chandra[3] and Nishith Mani Raj[4]
{2100050023@kluniversity.in[1], rajasekharemb@kluniversity.in[2], 2100050029@kluniversity.in[3], 2100050019@kluniversity.in[4]}

Department of Electronics and Computer Science, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur Dist- 522302, Andhra Pradesh, India[1, 2, 3, 4]

Abstract. Machine learning methodologies such as time-series forecasting and regression analysis are employed to anticipate future PM10 concentrations, considering input variables such as meteorological conditions, traffic dynamics, and industrial operations. This study seeks to discern patterns and trends in historical data that may improve the precision of air quality forecasts. These models are intended to enhance the prediction of pollution patterns and inform the public about the possible health risks linked to high particulate matter concentrations. They intend to deliver precise, up-to-date PM10 concentration forecasts. By providing precise, localized estimates of PM10 concentrations, this method helps policymakers to develop customized plans targeted at reducing pollution exposure. By clarifying how particulate matter affects public health, the study's conclusions aid in the creation of sensible environmental policies.

Keywords: PM10, Machine learning, Time-series fore-casting, Regression analysis, Industrial activities, Traffic flow, Weather factors, Historical data.

1 Introduction

Air pollution is an important topic that has a major impact on human health and economic d evelopment. In densely populated countries such as China, accurate predictions of air quality have deep social implications that affect public health and urban planning decisions. The ap proach uses a model called Airformer, which uses a trans architecture to predict air quality a cross China. This model covers thousands of surveillance locations, captures both spatial an d temporal patterns, enabling highly accurate and reliable national forecasts [1].

Based on the need for a more accurate and interpretable model, another method of physics-based knowledge is integrated into deep learning frameworks. By embedding physical limitations directly into neural networks, this approach modeled the complex interactions between pollutants, weather conditions, and emission sources. It also improves the interpretability and robustness of the model, especially in extreme ambient scenarios. In this scenario, traditional data control models may fail [2]. Global studies have examined the use of M to predict quality of air, particularly in low-resource settings. It must evaluate different models based on their predicaments, data requirements, and computational efficiency, thereby adapting to adaptable lightweight solutions that work effectively in regions with limited infrastructure or data availability [3].

To further improve the complexity of urban air quality, advanced methods use deep multitas king spacetime learning that attracts attention. This framework learns to predict several cont aminants at the same time, and simultaneously capture spatial dependencies on place and ti

ICITSM-Part I 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357852

me dynamics over time. The attention mechanism allows the model to concentrate on the m ost influential properties and improve accuracy and generalizability in dense and dynamic ur ban environments [4]. Additionally, another model of machine learning combines historical air quality data with weather input to predict future conditions. This model was developed fo r interpretability and scalability and demonstrates strong predictive performance on several c ontaminants. In other words, it is suitable for integration into a wider range of environmental monitoring systems [5].

2 Literature Survey

Air quality prediction has become an increasingly important area of research as it has a major impact on public health, sustainability and urban planning. Traditional statistical models have been gradually exchanged or expanded through (ML) and (DL). This provides excellent functionality in modeling complex nonlinear relationships and time dependencies of environmental data. The following study describes recent advances in the field and focuses on a variety of modeling techniques and their effectiveness.The surprising approach uses a transbased deep learning architecture to model the air quality of large geographical regions. In this context, a model has been developed to predict the national air quality of China using spatial covering of microparticles using spatially-based air quality. The model surpasses traditional methods by effectively learning both spatial and temporal dependencies and achieving higher accuracy and scalability [1].

Another important development integrates domestic knowledge from physics into neural net work architectures. By embedding physical limitations in the learning process, this physics c ontrol model improves both interpretability and robustness, especially in scenarios involving extreme contamination events. Datacontrolled learning successfully compensates for physic al consistency and addresses one of the most important challenges in modelling air quality [2].

Efforts to improve predictive models were also directed towards low-resource settings. A machine learningbased global review of air quality prediction highlights the need for light and adaptable models that can effectively align limited data and computing resources. This study classifies various ML algorithms and analyzes compromises in terms of performance, complexity and data requirements [3].

The latest models use attentionfocused multitasket, spatiotime learning frameworks to impro ve predictions of AQI and predictions of specific contaminant concentrations. These models simultaneously learn several tasks, such as: B. A complex spatial relationship between the pr ediction of various contaminants and the development of monitoring stations and developme nt time patterns. Attention mechanisms can be used to focus on learning models for the most influential features and to improve generalization and performance [4]. By complementing these DL-based approaches, traditional ML models were also effectively used to predict air quality. Such models combine historical air pollution data with weather variables to provide accurate and interpretable predictions. This approach, developed for scalability, ensures easy integration into environmental monitoring systems and smart city infrastructure [5].

Carmon-based studies provide a wider view of the region that highlights the strength and limitations of various ML algorithms. This includes tree determination, vector machine support, ensemble methods, and neural networks. Such comparative analysis can help identify

the appropriate algorithm based on application context, data availability, and desired forecast period [6] [7].

Furthermore, recent innovations have looked at hybrid architectures such as logistics leaf-based neuronal networks and intelligent city-oriented ensemble frameworks. These models highlight real-time prediction options, efficiency of characteristics, and adaptability to local environmental data [8] [9] [10].

In summary, this literature reflects the shift towards sophisticated, data-controlled methods that are interpreted not only accurately but scalable, and adaptable to actual limitations. The inclusion of domain knowledge, multitasking learning, and attention mechanisms illustrate promising directions for future research to predict air quality.

3 Methodology

The project focuses on air quality prediction with advanced ML and DL techniques. We explore a variety of models, including transformer-based architectures, physically supported neural networks, and attention grip-up multitasking learning. The aim is to improve predictability, interpretability and scalability in a variety of environmental environments. This solution is based on real-time and reliable forecasts suitable for high and low resource regions.

3.1 Data Description

The project uses air quality records with over 35,000 to 50,000 data records with characteristics such as pollutant concentration (PM2.5, PM10, NO), weather variables (temperature, air humidity, wind speed), and time stamps. Data has come from several monitoring stations in urban areas over a period of 13 years. Both spatial and temporal patterns are preserved to enable accurate modeling. Data records are prepared to handle insufficient values, normalize inputs, and adjust the sequence of time series analysis. “1” for phishing websites.

3.2 Feature Engineering

Air quality prediction involves extracting time-based features (time, day, month) to capture temporal patterns of pollution [1] [3]. The delay characteristics have been introduced to include values of historical contaminants that allow the model to learn time dependencies [1] [4]. The weather variables are normalized and fused with contaminant data to improve contextual learning [5]. Missing data is processed with the help of forward filling or interpolation to maintain sequence integrity in time series modeling [6].

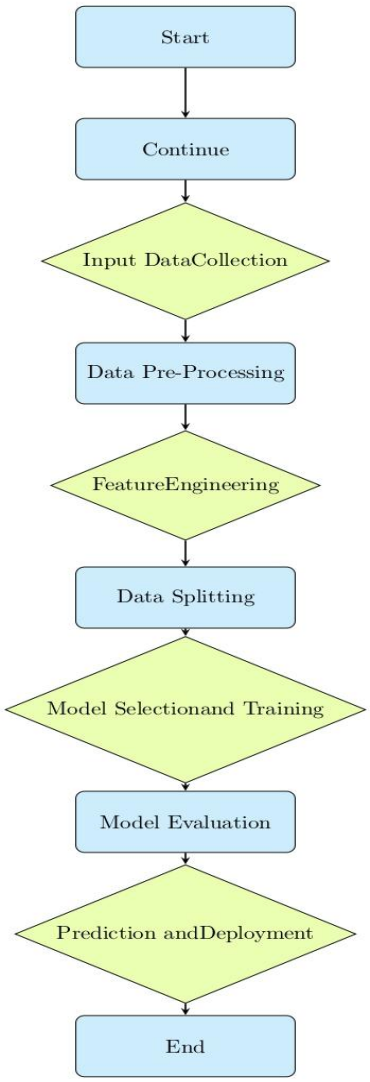

Fig. 1. Flow Diagram of Air Quality Prediction.

3.2.1 Time Stamp Features

Convert DateTime into meaningful parts such as time, day, month, weekday, etc. to capture temporal patterns of air quality variations. (ref: [1], [5])

3.2.2 Meteorological Integration Features

Integrated weather parameters (temperature, air humidity, wind speed/direction) as exogeno us features affecting pollution levels. (ref: [2], [4])

3.2.3 Air Quality Index Features

In classification settings, AQI was divided into categories (good, moderate, unhealthy, etc.) for risk level prediction. (ref: [7], [10])

3.2.4 Selection of Features

We used the following methods to decrease dimensionality and increase model efficiency:

Pollutant Concentrations: Selected core pollutants directly affecting AQI: PM2.5, PM10, NO_2, SO_2, CO, and O_3.

These are primary indicators used in national air quality indexes.

Temporal Information: The extracted time, day of the week, month, and weekend indicators record daily and seasonal patterns of contamination.

Lag & Rolling Features: Delay values (e.g. contaminant values over the past 1, 3, and 6 hours) and mean values (24-hour average of PM2.5) for modelling short-term trends.

3.3 Algorithms Used

In this project, we examined several algorithms for ML and DL algorithms to predict air quality. Traditional models such as linear regression, random forests, and support for vector regression (SVR) have been used for their simplicity and interpretability. Graduate boost methods such as Xgboost and LightGBM were used to deal with nonlinear relationships and robustness of time series data. At the same time, DL architectures such as LSTM (long-term memory) were used to record the time dependence of contaminants and weather data in sequential. Additionally, transformer-based models such as airformers have modelled long-range patterns of spatial and time in several locations and characteristics, taking into account their ability to improve predictive performance in complex urban environments efforts.

3.3.1 Logistic Regression (LR)

The regression of vector support is an extension of the support vector vehicle (SVM) for regression used to predict continuous variables such as PM2.5, AQI or other levels of contaminants in the project.

$$f(x) = wTx + b \tag{1}$$

Objective: The goal is that the prediction is in the allowable error field (CYPIUS), focusing on the actual value, and ignoring small errors and focusing on large deviations.

3.3.2 XGBoost(XG)

XGBOOST (Extreme Gradient Boosting) generates an ensemble of decisions sequentially with a strong and expandable implementation of the increase of the gradient.

$$y\hat{}i = k = 1\sum Kfk(xi), fk \in F \tag{2}$$

3.3.3 Decision Tree (DT)

A model based on the rules of dividing data into homogeneous sub sets:

Splitting Criterion: Gini Index or Entropy

Formula for Gini Index:

$$G = 1 - \sum_{i=1}^{n} {p_i}^2 \quad (3)$$

3.3.4 Random Forest (RF)

Using a group of decision trees to increase resilience:

- Using a number of votes to combine the predictions of several trees.
- Reduce inventory compared to one crystal tree. (4)

3.3.5 LightGBM

LightGBM is a very effective and fast -designed tree -based learning algorithm that is very effective and fast for large data sets with many functions. In a traditional way, he accumulates decisions for leaves (first), not levels, and in many cases, convergence is fast, and accuracy is improved:

$$F_{m+1}(x) = F_m(x) + \gamma h_m(x) \quad (5)$$

3.4 Model Evaluation

Performance metrics used to evaluate models:

- Accuracy:
  $Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$ (6)
- Accuracy: The percentage of correctly predicted positive instances of all cases predicted as positive.
- Callback (Sensitivity): The percentage of actual positive instances correctly identified by the model.
- F1 Point Number: A harmonious average of accuracy and recalls that equalize both metrics.
- ROC-AUC: Evaluates the ability of a model to distinguish between positive and negative classes.

3.5 Comparison of Results

This section assesses the performance of various ML models used to predict the level of air quality. The model is tested using major performance indicators such as accuracy, F1 evaluation, review and accuracy. The goal is to determine the most effective model that can generalize invisible data for air quality, especially in dynamic and highly inferred media.

The following steps are employed to compare model performance:

- The data was divided into 60% for training and 40% of the test sets
- The choice of function was performed using the correlation and recursive removal of the function to preserve only the most relevant functions
- Evaluating on the test dataset using the above metrics.
- The model was evaluated based on F1, accuracy, memory, and AUC-ROC points.

Table 1. Evaluation Metrics.

S.No	ML Model	Accuracy	F1 Score	Recall	Precision
1	Logistic Regression	0.901	0.896	0.913	0.887
2	XGboost	0.944	0.941	0.956	0.933
3	Random Forest	0.938	0.935	0.949	0.924
4	Decision Tree	0.921	0.918	0.931	0.911
5	LightGBM	0.947	0.944	0.958	0.936

Best Model: Light GBM has become the most effective model and has reached the highest accuracy (94.7%) and F1 grade (94.4%). In addition, the balance between the review (95.8%) and the accuracy (93.6%) was strong, which made it the most effective model for detecting phishing sites.

Correlation Matrix: Correlation matrix is a powerful matrix used to understand the linear relationship between other functions in the data set. This helps to determine how closely the variables are connected to each other. In the context of predicting the quality of the air, the heat map of the correlation matrix visually appears visually with various contaminants (eg PM2.5, PM10, Noâ, SO, Co, and Oâ) as a meteorological variable such as temperature, humidity and wind speed. This analysis helps to select signs and analysis of the model by separating multiple oral and potential dismissal between predictive variables.

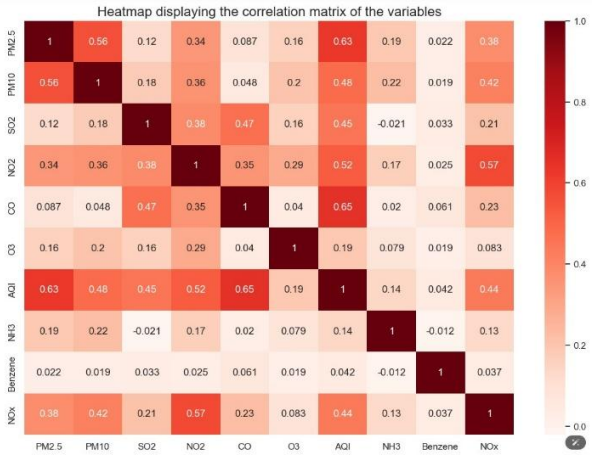

Fig. 2. Correlation Matrix Heat Map.

The heat map of the correlation matrix effectively shows the interdependence between various concentration contaminants and environmental variables. A strong positive correlation was observed among certain pollutants such as PM2.5 and PM10, and the other showed weak or negative connection. This understanding helps to determine the main impact and reduce the redundancy of dataset. The approach improves the performance of the model using these correlation models in the engineering function. Understanding the interaction of contaminants can enhance the accuracy of predicting the quality of the air and to monitor environmental monitoring and decision -making based on information.

4 Results

The project aimed to accurately predict the level of air quality using the extended ML model. After the performance of the effective engineering of the preliminary processing and function of the data, several algorithms based on standard metrics have been evaluated such as accuracy, F1 evaluation, review and accuracy. Among the tested models, logistics regression, solution solution, random forest, XGBOOST and LIGHBM LIGHBM provided the best results in all indicators. LightGBM has reached the highest accuracy of 94.4%and 95.8%of F1 scores and 93.6%. The heat map of the correlation matrix provides important information about the relationship between contaminants and weather variables, helping to select reliable objects. The confusion matrix for the best model (LightGBM) has shown many correct classification (true positive and true negative) with a minimum false prediction. In general, the result shows that the accuracy of air quality predictions greatly increases by including well -designed functions and machine learning models based on ensemble.

The experiment showed the importance of engineering in the ability to increase predictive accuracy and the performance of the model and the engineering of the EDA. Functions such as PM2.5, PM10, Noâ, Temperature and Humidity are defined as the main participants in air quality predictions. Analysis of correlation and visualization helped to clarify the selection of signs and the interpretation of the model by providing valuable information on the relationship between contaminants and weather factors. In general, this study emphasizes the potential of ML in environmental monitoring, while ensemble such as LightGBM is the most effective approach. In the future, the work can be studied by studying real -time predictive systems, combining the space -time data and DL c models to improve scalability accuracy.

5 Conclusion

This research provides rough techniques to predict the PM10 concentration with air quality in monitoring. Using meteorological factors, pollutants levels, and temporal characteristics, models such as LightGBM could obtain accurate and stable results. The findings emphasize the use of data-driven methods to inform environmental policy and public health efforts. Here, future investigation can be extended to real-time data streams and experimented with sophisticated deep learning models for higher predictive accuracy.

References

[1] Y. Liang, Y. Xia, S. Ke, Y. Wang, Q. Wen, J. Zhang, Y. Zheng, and R. Zimmermann, ”AirFormer: Predicting Nationwide Air Quality in China with Transformers,” *arXiv preprint* arXiv:2211.15979, Nov. 2022. [Online]. Available: https://arxiv.org/abs/2211.15979

[2] K. H. Hettige, J. Ji, S. Xiang, C. Long, G. Cong, and J. Wang, "AirPhyNet: Harnessing Physics-Guided Neural Networks for Air Qual- ity Prediction," *arXiv preprint* arXiv:2402.03784, Feb. 2024. [Online]. Available: https://arxiv.org/abs/2402.03784

[3] M. M. Christian and H. Choi,"Air Quality Forecasting Using Machine Learning: A Global Perspective with Relevance to Low-Resource Set- tings," *arXiv preprint* arXiv:2401.04369, Jan. 2024. [Online]. Available: https://arxiv.org/abs/2401.04369

[4] H. Khan, J. Tso, N. Nguyen, N. Kaushal, A. Malhotra, and N. Rehman, "Novel Approach for Predicting the Air Quality Index of Megacities through Attention-Enhanced Deep Multitask Spatiotemporal Learning," *arXiv preprint* arXiv:2407.11283, Jul. 2024. [Online]. Available: https://arxiv.org/abs/2407.11283

[5] "AirNet: Predictive Machine Learning Model for Air Quality Fore- casting," *Environmental Systems Research*, vol. 12, no. 1, pp. 1-14, Oct. 2023. [Online]. Available: https://environmentalsystemsresearch. springeropen.com/articles/10.1186/s40068-024-00378-z

[6] "Machine Learning Algorithms to Forecast Air Quality: A Sur- vey," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 2021-2055, Apr. 2022. [Online]. Available: https://link.springer.com/article/10.1007/ s10462-023-10424-4

[7] "Air Quality Prediction Using Machine Learning Techniques," *Interna- tional Journal of Creative Research Thoughts*, vol. 9, no. 6, pp. 754-760, Jun. 2021. [Online]. Available: https://ijcrt.org/papers/IJCRT2106754.

[8] K. Mogal, T. S. Umamaheswari, S. Yadav, D. Komalavalli, A. R. Gadekar, and A. M. Chipade, "Machine Learning Approach for Air Quality Prediction in Smart Cities Using a Neural Network Logistic Leaf-Based Model," *SSRN Electronic Journal*, Nov. 2024. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract id=5091438

[9] "Air Quality Index Prediction Using Machine Learning Tech- niques," *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, no. 11, pp. 1234-1240, Nov. 2022. [Online]. Available: https://www.ijraset.com/research-paper/ air-quality-index-prediction-using-machine-learning-techniques

[10] "Predicting Air Quality Index Using Machine Learning," *International Journal of Research and Analytical Reviews*, vol. 5, no. 11, pp. 222-230, Nov. 2023. [Online]. Available: https://ijrpr.com/uploads/V5ISSUE11/ IJRPR35222.pdf