Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 3, March : 2025

Enhancing Air Pollution Forecasting with Machine Learning: A Comprehensive Analysis

M DHARMA VARDHANI, DEPARTMENT OF COMPUTER SCIENCE, Sri Durga Malleswara Siddhartha Mahila Klalasla, Vijayawada, AP, India.

Abstract:

Air pollution is a persistent global concern, affecting human health and ecosystems. Accurate forecasting is crucial for timely interventions and policy implementation. This study explores the application of machine learning techniques for air quality prediction using historical and real-time data. Various models, including regression, decision trees, and neural networks, are analyzed for their predictive capabilities. The findings demonstrate the effectiveness of machine learning in forecasting air pollution levels, contributing to environmental monitoring and mitigation strategies. In some of air pollution and air pollution is direct impact on human body. As we know that major pollutants are arising from Nitrogen Oxide, Carbon Monoxide & Particulate matter (PM), SO2 etc. Carbon Monoxide is arising due to the deficient Oxidization of propellant like as petroleum, gas, etc. nitrogen oxide (NO) is arising due to the ignition of thermal fuel; Sulphur Dioxide (So2) is major spread in air, So2 is a gas which is present more pollutants in air, it's affect more in human body. The predominance of air is overstated by multidimensional impacts containing spot, time and vague boundaries. Thegoal of this improvement is to take a gander at the AI basically based ways for air quality expectation. In this paper we will predict of air pollution by using machine learning algorithm.

Keywords: Air pollution forecasting, Machine learning, Environmental analysis, Neural networks, Predictive modeling

Introduction

The Environment describe about the thing which is everything happening in encircles the Environment is polluted by human daily activities which include like air pollution, noise pollution. If humidity is increasing more than automatically environment is going more hotter. Major cause of increasing pollution is increasing day by day transport and industries there are 75 % NO or other gas like CO, SO2 and other particle is exist in environment.. The expanding scene, vehicles and creations square measure harming all the air at a feared rate.

Therefore, we have taken some attributes data like vehicles no., Pollutants attributes for prediction of pollution in specific zone of

Delhi.The Environment describe about the thing which is everything happening in encircles the Environment is polluted by human daily activities which include like air pollution, noise pollution. If humidity is increasing more than automatically environment is going more hotter. Major cause of increasing pollution is increasing day by day transport and industries there are 75 % NO or other gas like CO, SO2 and other particle is exist in environment.. The expanding scene, vehicles and creations square measure harming all the air at a feared rate.

Therefore, we have taken some attributes data like vehicles no., Pollutants attributes for prediction of pollution in specific zone of Delhi.

Literature Survey:

Ni, X.Y.; Huang, H.; Du, W.P. "Relevance analysis and short-term prediction of PM

2.5 concentrations in Beijing based on multi-source data." Atmos. Environ. 2017, 150, 146-161.

The $PM_{2.5}$ problem is proving to be a major public crisis and is of great public-concern requiring an urgent response. Information about, and prediction of $PM_{2.5}$ from the perspective of atmospheric dynamic theory is still limited due to the complexity of the formation and development of $PM_{2.5}$. In this paper, we attempted to realize the relevance analysis and short- term prediction of $PM_{2.5}$ concentrations in Beijing, China, using multi-source data mining. A correlation analysis model of $PM_{2.5}$ to physical data (meteorological data, including regional average rainfall, daily mean temperature, average relative humidity, average wind speed, maximum wind speed, and other pollutant concentration data, including CO, NO_2, SO_2, PM_{10}) and social media data (microblog data) was proposed, based on the Multivariate Statistical Analysis method. The study found that during these factors, the value of average wind speed, the concentrations of CO, NO_2, PM_{10}, and the daily number of microblog entries with key words _Beijing; Air pollution' show high mathematical correlation with $PM_{2.5}$ concentrations. The correlation analysis was further studied based on a big data's machine learning model-Back Propagation Neural Network (hereinafter referred to as BPNN) model. It was found that the BPNN method performs better in correlation mining. Finally, an Autoregressive Integrated Moving Average (hereinafter referred to as ARIMA) Time Series model was applied in this paper to explore the prediction of $PM_{2.5}$ in the short- term time series. The predicted results were in good agreement with the observed data. This study is useful for helping realize real-time monitoring, analysis and pre-warning of $PM_{2.5}$ and it also helps to

broaden the application of big data and the multi-source data mining methods.

G. Corani and M. Scanagatta, "Air pollution prediction via multi-label classification," Environ. Model. Softw., vol. 80, pp. 259-264,2016.

A Bayesian network classifier can be used to estimate the probability of an air pollutant overcoming a certain threshold. Yet multiple predictions are typically required

regarding variables which are stochastically dependent, such as ozone measured in multiple stations or assessed according to by different indicators. The common practice (independent approach) is to devise an independent classifier for each class variable being predicted; yet this approach overlooks the dependencies among the class variables. By appropriately modeling such dependencies one can improve the accuracy of the forecasts. We address this problem by designing a multi-label classifier, which simultaneously predict multiple air pollution variables. To this end we design a multi-label classifier based on Bayesian networks and learn its structure through structural learning. We present experiments in three different case studies regarding the prediction of PM2.5 and ozone. The multi-label classifier outperforms the independent approach, allowing to take better decisions.

Mrs. A. GnanaSoundariMtech, (Phd) ,Mrs. J. GnanaJeslin M.E, (Phd), Akshaya

A.C. “Indian Air Quality Prediction And Analysis Using Machine Learning”. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019 (Special Issue).

Examining and protecting air quality has become one of the most essential activities for the government in many industrial and urban areas today. The meteorological and traffic factors, burning of fossil fuels, and industrial parameters play significant roles in air pollution. With this increasing air pollution, Weare in need of implementing models which will recordinformation about concentrations of air pollutants(so2,no2,etc).The deposition of this harmful gases in the air is affecting the quality of people‘s lives, especially in urban areas. Lately, many researchers began to use Big Data Analytics approach as there are environmental sensing networks and sensor data available. In this paper, machine learning techniques are used to predict the concentration of so2 in the environment. Sulphur dioxide irritates the skin and mucous membranes of the eyes, nose, throat, and lungs. Models in time series are employed to predict the so2 readings in nearing years or months

The Environment describe about the thing which is everything happening in encircles the

Environment is polluted by human daily activities which include like air pollution, noise pollution. If humidity is increasing more than automatically environment is going more hotter. Major cause of increasing pollution is increasing day by day transport and industries there are 75 % NO or other gas like CO, SO2 and other particle is exist in environment..

Introduction Air pollution causes severe health risks, including respiratory diseases and cardiovascular conditions, while also leading to environmental degradation. Traditional monitoring systems rely on physical sensors, which, despite their precision, are expensive and spatially limited. Machine learning provides an efficient alternative by analyzing past pollution data to develop predictive models, enabling proactive pollution management.

Existing Disadvantages:

High Cost of Traditional Systems: Air quality monitoring stations require significant investment in infrastructure and maintenance.

Geographical Limitations: Sensors are often sparsely distributed, leading to gaps in data coverage.Delayed Response: Conventional monitoring detects pollution after it has already impacted the environment and public health.

Data Processing Challenges: Large volumes of air pollution data demand efficient computational methods for real-time analysis.

Proposed System:

The proposed system integrates advanced machine learning algorithms with real-time data streaming to improve air pollution forecasting. The key components of this system include:Real-Time Data Integration: Continuous collection of air quality data from IoT-based sensors and satellite feeds.Hybrid Modeling Approach: Combining deep learning (LSTM, CNN) with traditional ML models for enhanced accuracy.Automated Feature Engineering: Utilizing AI-driven methods for selecting relevant features dynamically.

Cloud-Based Deployment: Hosting models on cloud platforms for scalability and real-time analytics.Early Warning Mechanism: Alerting authorities and the public about potential pollution spikes through mobile apps and dashboards.

Proposed Advantages:

Improved Predictive Accuracy: Advanced models, particularly deep learning, enhance forecasting precision.

Cost-Effective Alternative: Machine learning reduces dependency on expensive physical sensors.

Real-Time Forecasting: AI-driven models can predict pollution levels ahead of time, allowing preventive measures.

Scalability: ML models can be trained on diverse datasets and applied to various geographical regions.

Related Work

Several studies have employed statistical and deep learning techniques to predict air quality. Common approaches include regression models, support vector machines, and neural networks. Integrating meteorological and traffic data has been found to enhance model accuracy. However, challenges remain in achieving real-time prediction efficiency and handling large-scale datasets.

Methodology This research utilizes a dataset containing pollutant concentrations such as PM2.5, NO2, and SO2, along with meteorological factors like temperature, humidity, and wind speed. Data preprocessing methods, including normalization and feature selection, are applied before training the models.

Data Acquisition: Compilation of historical air quality data from monitoring stations.

Feature Selection: Identification of crucial variables affecting pollution levels.

Model Implementation: Evaluation of Linear Regression, Random Forest, and LSTM neural networks.

Training & Validation: Assessment of model performance using RMSE and R-squared metrics.

Decision tree classifiers

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with this class.

Step 2. Otherwise, let T be some test with possible outcomes O1, O2,…, On. Each object in Shas one outcome for T so the test partitions S into subsets S1, S2,… Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

Gradient boosting

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 3, March : 2025

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.[1][2] When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes

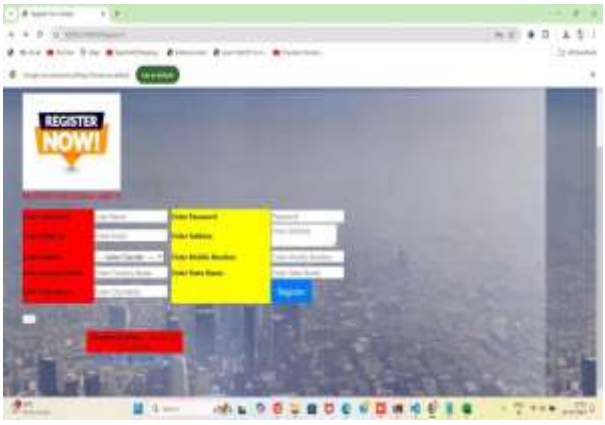
the other methods by allowing optimization of an arbitrary differentiable loss function.

K-Nearest Neighbors (KNN)

Simple, but a very powerful classification algorithmClassifies based on a similarity measure Non-parametric Lazy learning Does not –learn‖ until the test example is given Whenever we have a new data to classify, we find its K-nearest neighbors from thetraining data

Results and Discussion

A comparative analysis of predictive models indicates that LSTM neural networks outperform traditional regression techniques in terms of accuracy. While Random Forest provides reliable short-term predictions, deep learning models demonstrate superior adaptability to fluctuating environmental conditions. The integration of real-time data sources can further enhance predictive reliability.



Conclusion

Machine learning is proving to be a transformative tool in air pollution forecasting, offering data-driven insights that aid urban planning and public health initiatives. By leveraging historical and real-time environmental data, ML models can predict pollution levels with higher accuracy than conventional methods, enabling timely interventions. Future advancements should focus on integrating multiple real-time sensor data sources to enhance model precision and adaptability. Additionally, the development of deep learning frameworks—such as hybrid neural networks that combine convolutional and recurrent architectures—can further refine dynamic pollution predictions, allowing for more effective mitigation strategies and policy implementations.

REFERENCES:

- Ni, X.Y.; Huang, H.; Du, W.P. –Relevance analysis and short-term prediction of PM 2.5 concentrations in Beijing based on multi-source data.‖ Atmos. Environ. 2017, 150, 146-161.
- G. Corani and M. Scanagatta, "Air pollution prediction via multi-label classification," Environ. Model. Softw., vol. 80, pp. 259-264,2016.
- Mrs. A. GnanaSoundariMtech, (Phd) ,Mrs. J. GnanaJeslin M.E, (Phd), Akshaya A.C. –Indian Air Quality Prediction And Analysis Using Machine Learning‖. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019 (Special Issue).
- Suhasini V. Kottur , Dr. S. S. Mantha. –An Integrated Model Using Artificial Neural Network
- RuchiRaturi, Dr. J.R. Prasad .―Recognition Of Future Air Quality Index Using Artificial Neural Network‖.International Research Journal ofEngineering and Technology (IRJET) .e-ISSN: 2395-0056 p-ISSN: 2395-0072 Volume: 05 Issue: 03 Mar-2018
- Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu .‖ Detection and Prediction of Air Pollution using Machine Learning Models‖. International Journal o f Engineering Trends and Technology (IJETT) - volume 59 Issue 4 - May 2018
- Gaganjot Kaur Kang, Jerry ZeyuGao, Sen Chiao, Shengqiang Lu, and Gang Xie.‖ Air Quality Prediction: Big Data and Machine Learning Approaches‖. International Journal Environmental Science and Development, Vol. 9, No. 1, January 2018

- PING-WEI SOH, JIA-WEI CHANG, AND JEN-WEI HUANG,‖ Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations,‖ IEEE ACCESSJuly 30, 2018.Digital Object Identifier10.1109/ACCESS.2018.2849820.
- GaganjotKaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie,‖Air Quality Prediction: Big Data and Machine Learning Approaches,‖ International Journal of Environmental Science and Development, Vol. 9, No. 1, January2018.