WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Air Quality Prediction Using Prediction Algorithms

Dhulipudi Varsha Sri Lakshmi

B.Tech Student, Department of IT,GMR Institute of Technology, Rajam-532127, Andhra Pradesh, India
Email: 21341A1233@gmrit.edu.in

ABSTRACT

Air quality has become a critical concern in urban environments due to its profound impact on public health and the environment. As machine learning is a type of artificial intelligence that can be used to predict the quality of air as it has gained prominence in recent years for their ability to predict and monitor air quality and its disadvantageous effects based on the historical data to identify patterns and relationships between air pollutants and other factors, the air pollutants include carbon monoxide, ozone, sulphur dioxide, nitrogen dioxide and so on. This abstract provides an overview of the algorithms and the methodologies used in machine learning for air quality prediction. This paper reviews various machine learning approaches, including regression models, hybrid models and the algorithms applied to air quality prediction. The paper concludes with an emphasis on the potential benefits of deploying machine learningbased air quality prediction systems for public health and policy decision making. It explains the need for ongoing research to develop more accurate and reliable models, promote real-time monitoring and facilitate proactive measures to improve air quality in urban areas.

Keywords: Air Quality Prediction, Regression Models, Urban environments, Prediction algorithms

INTRODUCTION

Air pollution stands as a pressing global challenge, posing significant threats to human health and the environment. Central to this issue is the concern surrounding PM2.5, fine particulate matter with a diameter of 2.5 micrometers or smaller, known for its adverse effects on respiratory health and its contribution to environmental degradation. In this context, the following five research articles converge on the urgent need to address air pollution, with a specific focus on predicting PM2.5 concentrations. These studies recognize the critical role of accurate air quality forecasting as an instrumental tool for effective pollution mitigation and control efforts.

Research Focus on PM2.5 Concentration: The collective emphasis of these research articles lies in the meticulous examination of PM2.5 concentration, recognizing it as a pivotal indicator of air pollution's severity. These fine particles, often originating from combustion processes and industrial activities, not only pose immediate health risks but also have long-term environmental implications. By honing in on PM2.5, these studies underscore the need for predictive models and tools that can offer insights into its concentration levels, aiding in the development of targeted strategies for pollution management and control.

Importance of Precise Air Quality Forecasting: The overarching theme within these articles' centers on the significance of precise air quality forecasting in the battle against air pollution. Accurate predictions of PM2.5 concentrations are identified as crucial for implementing timely interventions and formulating effective policies aimed at reducing pollutant levels. The collective assertion is that robust forecasting models not only enhance our understanding of pollution patterns but also empower decision-makers and environmental agencies to proactively address the health and environmental consequences associated with elevated PM2.5 levels. In essence, these research articles advocate for a proactive and predictive approach to air quality management to safeguard public health and the environment in the face of escalating air pollution challenges.

LITERATURE REVIEW

The examination of particulate matter (PM) concentration prediction in urban environments has been a focal point in several scholarly investigations, concentrating on various regions including Curitiba, Brazil, Indian cities, California, and Taiwan. These studies collectively employed a range of machine learning (ML) algorithms to tackle the challenge of predicting air quality levels, predominantly focusing on PM10 and PM2.5 levels.

Model Utilization:

Random Forest for PM10 Prediction: Notably, in studies focused on Curitiba, Brazil, the Random Forest (RF) algorithm stood out as a prominent choice for predicting PM10 concentrations. RF, known for its ensemble nature, was preferred due to its capacity to handle multiple input variables derived from meteorological data and vehicle flow statistics. The model was trained using historical PM measurements from monitoring stations in Curitiba.

Comparative Study of ML Models: The investigation on Indian cities involved a comparative analysis of multiple ML models. Gaussian Naive Bayes, Support Vector Machine (SVM), XG Boost, and Random Forest were among the models considered for air quality prediction. The emphasis on employing diverse ML models was aimed at a comprehensive evaluation, recognizing the multifaceted nature of air pollution.

SVR for Precise Air Quality Prediction: Studies focusing on California delved into the use of Support Vector Regression (SVR) with a Radial Basis Function (RBF) kernel. SVR was deemed suitable for modeling the dynamic and variable nature of air pollutants. The anticipated results included precise hourly predictions of pollutant concentrations and Air Quality Index (AQI) for California, with a target accuracy of 94.1% in AQI classification.

Gradient Boosting for PM2.5 Forecasting: The exploration in Taiwan involved the application of Gradient Boosting Regressor, specifically targeting the forecasting of PM2.5 concentrations. The model utilized statistical metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) to evaluate the accuracy of predictions. However, limitations included a focus on a specific region and a reliance on historical data (2012-2017), potentially affecting the generalizability of findings.

Key Objectives and Expected Outcomes:

Health Implications and Need for Accurate Predictions: Across all studies, a common thread highlighted the detrimental effects of PM on human health, especially for vulnerable populations. Accurate air quality monitoring and prediction models were emphasized as essential for proactive measures to mitigate health risks.

Accuracy Targets and Suitability of Models: The anticipated outcomes across studies revolved around achieving high accuracy in air quality predictions, often targeting specific percentages, such as 80.42% and 94.1% accuracy. The choice of models was driven by their abilities to handle complex air pollution dynamics, such as seasonal variations and pollutant sources.

Challenges and Future Considerations:

Data Challenges and Generalizability: Challenges included data imbalance, seasonal variations impacting pollution levels, and limitations in generalizability beyond specific regions or datasets. Future research considerations revolved around exploring diverse datasets, hyperparameter optimization, and comparing different algorithms to enhance predictive accuracy.

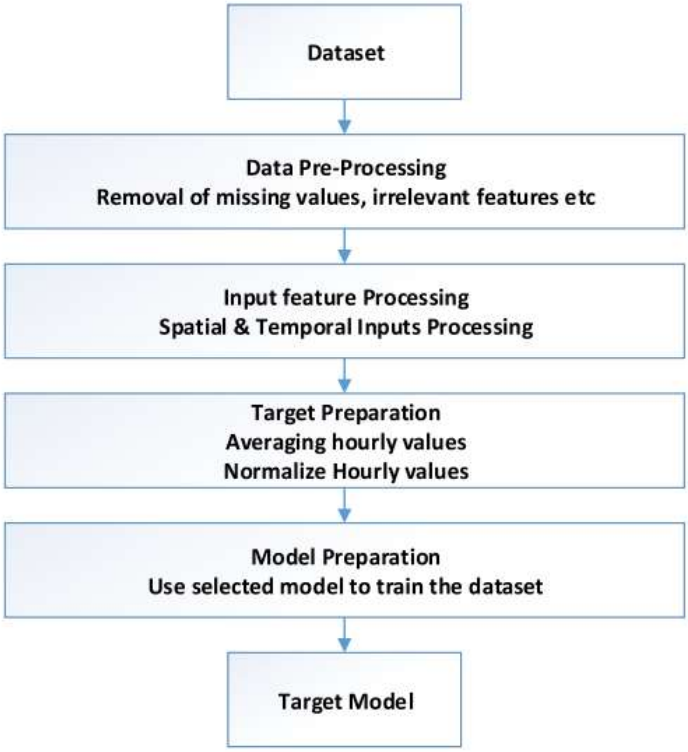
Algorithmic Limitations and Scope: Limitations were noted, such as a potential lack of generalizability of models beyond specific regions or datasets. The focus on accuracy and processing time was highlighted as potentially overlooking other crucial factors, like energy consumption and model interpretability.

METHODOLOGY:

1. Model Selection and Execution:

Machine learning, a pivotal tool for air quality prognosis, encompasses a spectrum of models, with Random Forest (RF) and Support Vector Regression (SVR) wielding prominence in recent studies. RF, known for its ensemble nature, assumed a leading role in predicting PM10 concentrations in Curitiba, Brazil. Its strength lies in the fusion of multiple decision trees, culminating in a robust framework adept at handling intricate, nonlinear associations embedded within the data. In contrast, SVR coupled with the Radial Basis Function (RBF) kernel, a standout in analyses spanning Indian cities and California, shines in capturing the intricate, nonlinear interplay between variables. Both models underwent rigorous training: RF assimilated historical PM data, meteorological factors, and urban traffic records, while SVR ingested hourly pollutant and meteorological datasets, ensuring comprehensive learning and prediction.

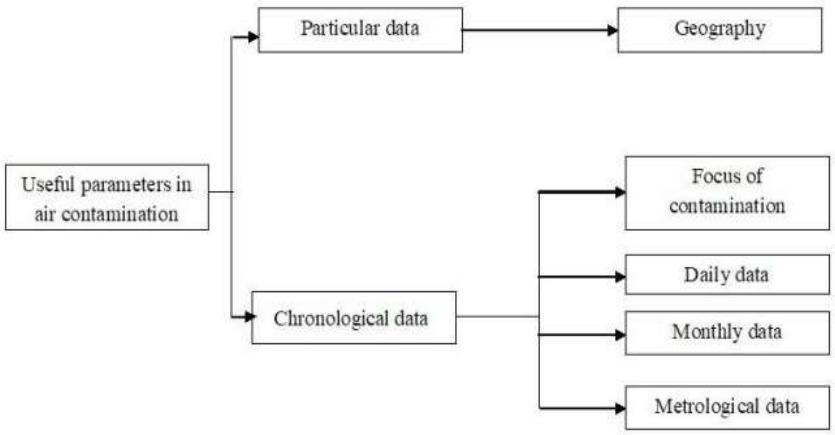
RF's prowess in Curitiba's study stemmed from its ability to synthesize diverse data types, harnessing historical PM data and various meteorological factors alongside urban traffic details. This amalgamation empowered the model to discern complex patterns within the data, crucial for accurate PM10 predictions. Conversely, SVR with RBF kernel exhibited its strength across varied geographies by effectively navigating intricate nonlinear relationships, grasping the nuances between pollutant concentrations, meteorological parameters, and air quality indices. The comprehensive learning approach in SVR, fed with hourly data on pollutants and meteorological insights, facilitated a robust understanding of intricate relationships within the datasets from Indian cities and California.



These models underwent meticulous training, leveraging historical records and real-time data to fortify their predictive capabilities. RF's ensemble of decision trees synergized historical PM concentrations, meteorological variables, and vehicular flow patterns to elucidate predictive patterns. SVR's utilization of hourly datasets for pollutants and meteorological factors fostered an in-depth understanding of temporal variations, enabling accurate forecasts across diverse regions. Both methodologies, though distinct in their approaches, converged in their pursuit of elucidating the complex dynamics underlying air quality, shedding light on the potential and challenges inherent in using machine learning for environmental prognosis.

2. Dataset Descriptions:

The datasets used in these studies showcased diversity, representing various geographical regions and focusing on distinct pollutants. Curitiba's dataset encompassed a historical record of PM measurements, meteorological insights, and urban traffic information, offering a comprehensive understanding of the interplay between various factors influencing air quality.



The Indian city dataset, derived from the Central Pollution Control Board (CPCB), spanned from 2015 to 2020, illuminating trends and patterns in air quality over time. California's dataset, sourced from the EPA's Air Quality records between 2016 and 2018, provided hourly readings of numerous pollutants and meteorological factors. Similarly, the Taiwan Air Quality Monitoring Network (TAQMN) dataset, spanning from 2012 to 2017, specifically focused on PM2.5 levels across selected cities.

3. Evaluation Metrics Utilized:

Assessment metrics encompassed multi-faceted dimensions, delving into various aspects influencing air quality prediction. These included evaluating the health impact of pollutants (Air Quality Metrics), understanding the influence of meteorological conditions on air quality dynamics (Meteorological Metrics), and unraveling the relationship between traffic density and pollution levels (Vehicle Flow Metrics). Additionally, these studies prioritized evaluating overall data quality to ensure the reliability and integrity of the datasets. Moreover, modeling and predictive metrics, such as R-squared, MAE, and RMSE, played a crucial role in scrutinizing the models' accuracy in predicting pollutant concentrations and AQI.

4. Key Findings and Validation:

The methodologies revealed insightful outcomes. For instance, the RF model in Curitiba showcased optimal performance at a daily time scale, meeting predefined accuracy targets. Conversely, SVR models employed in Indian cities and California demonstrated exceptional predictive capabilities, reaching a high accuracy of 94.1% in forecasting pollutant concentrations and AQI values. These findings underscored the models' proficiency in capturing intricate patterns within pollutant data, providing reliable predictions essential for environmental and public health assessments.

5. Data Handling and Limitations:

The methodologies employed meticulous data preprocessing steps, including normalization, handling missing values, and addressing outliers. Despite these efforts, acknowledged limitations included the study's focus on specific pollutants or cities, potential issues related to data imbalance, and the necessity for broader validation across diverse datasets and geographic locations. However, these methodologies laid a robust foundation, amalgamating model selection, data curation, evaluation metrics, and insightful findings, furthering the discourse on air quality prediction using machine learning.

RESULTS

The insights derived from the five studies underscore the versatility of machine learning models in predicting air quality parameters, notably particulate matter concentrations like PM10 and PM2.5. Each study employed different methodologies and models, highlighting their unique strengths and performance in distinct contexts.

Prediction of Particulate Matter Concentration in Urban Environment using Random Forest: This study focused on Curitiba, Brazil, employing RF models to forecast PM10 concentrations. It identified meteorological variables and vehicle flow as key predictors. The correlation between vehicular emissions and PM10 levels was evident, highlighting the impact of traffic. The study emphasized the significance of incorporating baseline PM10 concentrations for higher prediction accuracy.

Air Pollution Prediction with Machine Learning: A Case Study of Indian Cities: This research centered on Indian cities, using ML models to predict AQI. Notably, XG Boost exhibited superior accuracy in predicting AQI levels. It addressed data imbalances and highlighted seasonal patterns in pollutants like PM2.5 and PM10, noting their varying levels over time.

A Machine Learning Approach to Predict Air Quality in California: Focused on California, this study employed SVR to forecast pollutant levels and AQI. It emphasized SVR's accuracy in modelling pollutants like O3, CO, and SO2, proposing future enhancements by enriching datasets and optimizing SVR parameters for better precision.

Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities: Examined pollution levels across cities in China using regression techniques. Random Forest consistently displayed robust accuracy in predicting pollution levels, showcasing its efficiency in forecasting based on extensive historical data.

Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models: Focused on Taiwan's major cities, comparing various ML models to forecast PM2.5 levels. The Gradient Boosting Regressor emerged as the most accurate, demonstrating high accuracy in forecasting PM2.5 levels in both training and testing datasets.

Overall, the studies collectively underscore the potential of machine learning models, such as Random Forest, XG Boost, SVR, and Gradient Boosting, in forecasting air quality parameters, offering insights crucial for environmental monitoring and public health assessments.

CONCLUSION

The analysis and discussions across these five research papers underscore the complexity and significance of predicting air quality and particulate matter concentrations in various urban environments. The findings reveal the efficacy of different machine learning techniques in forecasting pollution levels and emphasize crucial factors influencing these predictions. Firstly, the studies highlighted the significance of meteorological variables, vehicular emissions, and sensor calibration in accurately predicting PM10 and PM2.5 concentrations. They elucidated the correlations between pollutant levels and factors like temperature, wind speed, solar radiation, and vehicle flow, underscoring their impact on air quality. Additionally, the investigation into machine learning models, including Gradient Boosting, Random Forest, and others, showcased their varied performances across different scenarios and datasets. Techniques like Gradient Boosting exhibited exceptional accuracy, especially in forecasting PM2.5 levels, affirming their potential in air quality prediction. Moreover, the studies proposed avenues for future research, emphasizing dataset enrichment, parameter optimization in models like support vector regression, and comparative analyses with diverse machine learning algorithms. These suggestions aim to improve the precision and reliability of

air quality forecasting models.In conclusion, these research endeavors collectively contribute comprehensive insights into the dynamics of air quality prediction, showcasing the potential of machine learning techniques while outlining critical areas for further exploration and refinement. Overall, they provide a foundation for developing more accurate and robust models for forecasting air pollution levels in urban environments

References

[1]. Emilio Graciliano Ferreira Mercuri, Isadora Bergami, Steffen Manfred Noe, Heikki Junninen, and Ulrich Norbisrath. 2023. Prediction of particulate matter concentration in urban environment using Random Forest.

[2]. Kumar, K., & Pande, B. P. (2023). Air Pollution Prediction with Machine Learning: A Case Study of Indian Cities.

[3]. Castelli, M., Clemente, F. M., Popovicˇ, A., Silva, S., & Vanneschi, L. (2020). A Machine Learning Approach to Predict Air Quality in California.

[4]. Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., & Asghar, M. N.(2019). Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities.

[5]. Doreswamy, H. K. S., Yogesh, K. M., & Gad, I. (2023). Forecasting Air PollutionParticulate Matter (PM2.5) Using Machine Learning Regression Models.

[6]. Instituto Brasileiro de Geografia e Estatística IBGE. 2022. Sinopse do Censo Demográfico 2022. Instituto Brasileiro de Geografia e Estatística - IBGE, Rio de Janeiro.https://cidades.ibge.gov.br/brasil/pr/curitiba/pesquisa/22/28120

[7]. Saverio De Vito, Elena Esposito, Nuria Castell, Philipp Schneider, and Alena Bartonova.2020. On the robustness of field calibration for smart air quality monitors. Sensors and Actuators B: Chemical 310 (2020), 127869

[8]. Martha Arbayani Zaidan, Naser Hossein Motlagh, Pak Lun Fung, Abedalaziz S Khalaf, Yutaka Matsumi, Aijun Ding, Sasu Tarkoma, Tuukka Petäjä, Markku Kulmala, and Tareq Hussein.(2022)Intelligent air pollution sensors calibration for extreme events and drifts monitoring. IEEE Transactions on Industrial Informatics19, 2 (2022), 1366–1379

[9]. Wan Nur Shaziayani, Ahmad Zia Ul-Saufie, Sofianita Mutalib, Norazian Mo hamad Noor, And Nazatul Syadia Zainordin. 2022. Classification Prediction of PM10 Concentration Using a Tree Based Machine Learning Approach. Atmo sphere 13, 4 (2022), 538.

[10]. Departamento Nacional de Trânsito (Denatran). Ministério da Infraestrutura. 2022. Frota de Veículos - 2022. https://www.gov.br/infraestrutura/pt-br/assuntos/ transito/conteudo-Senatran/frota-de-veiculos-2022. Accessed: 2023-04-05.