Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 5 Issue 3 (2025)

Predictive modeling of air pollution levels: A state-of-the-art review of machine learning techniques

Ms. Reena G.Bhati[1], Mr. Mayur Dutta[2]
[1,2]Assistant professor, tilak maharashtra vidyapeeth, pune
[1]reena4bhati@gmail.com and [2]duttamayur888@gmail.com

Abstract
Air pollution remains a significant environmental concern, necessitating accurate monitoring and forecasting techniques. This survey paper reviews state-of-the-art machine learning approaches for air quality prediction, including artificial intelligence, decision trees, deep learning, and ensemble methods. It examines data sources, preprocessing techniques, and the core algorithms employed across different pollutants and regions. The primary objective of this survey paper is to conduct a comprehensive examination of various big data analytics and machine learning methodologies that have been employed for the purpose of forecasting air quality levels. The paper provides an in-depth review and synthesis of existing published research studies that have utilized artificial intelligence techniques, decision tree algorithms, deep learning models, and other advanced approaches to evaluate and predict air quality indicators. Additionally, the survey sheds light on the current challenges faced in this domain and identifies potential areas that necessitate further investigation and research efforts.

Keywrods: Air quality evaluation, big data analytics, data-driven air quality evaluation, and air quality prediction.

Introduction

Air pollution has emerged as a significant global concern, posing severe risks to human health, ecosystems, and environmental sustainability. The rapid urbanization, industrialization, and overreliance on fossil fuels have led to the accumulation of harmful pollutants in the atmosphere, such as particulate matter (pm), nitrogen oxides (nox), sulfur dioxide (so2), and ground-level ozone (o3) [1]. Exposure to these pollutants has been linked to various respiratory diseases, cardiovascular problems, and even premature mortality [2]. Consequently, air quality monitoring and forecasting have become crucial for implementing effective mitigation strategies, informing the public, and enabling data-driven policymaking.

Traditional approaches to air quality monitoring relied heavily on ground-based monitoring stations, which provided localized measurements but often lacked the spatial and temporal resolution required for comprehensive assessment [3]. The advent of advanced sensing technologies, such as satellite observations and low-cost sensor networks, has revolutionized the field by enabling the collection of vast amounts of environmental data [4]. However, the sheer volume, velocity, and variety of this data pose significant challenges for conventional analytical techniques.

This is where machine learning (ml) and big data analytics have emerged as powerful tools for air quality prediction [5]. Ml algorithms can effectively capture the complex, non-linear relationships between air pollutants and various influencing factors, such as meteorological conditions, land use patterns, and emission sources [6]. Moreover, the ability of ml models to learn from historical data and continuously adapt to new observations makes them well-suited for real-time air quality forecasting [7].

Over the past decade, researchers have explored a wide range of ml techniques for air quality prediction, including traditional methods like linear regression, decision trees, and ensemble models [8],

as well as more advanced approaches like artificial neural networks (anns) [9], deep learning [10], and hybrid models that combine ml with physical models [11]. These techniques have been applied to forecast various air pollutants, such as pm2.5 [12], pm10 [13], no2 [14], and o3 [15], across diverse geographical regions and urban environments.

Despite the promising results, several challenges persist in the application of ml for air quality prediction. These include the availability and quality of training data [16], the complexity of atmospheric processes [17], and the interpretability and generalizability of ml models [18]. Additionally, the integration of multi-source data, such as satellite observations, ground-based measurements, and meteorological data, presents opportunities for improving prediction accuracy [19].

This survey paper aims to provide a comprehensive review of the state-of-the-art ml techniques for air quality forecasting, covering a wide range of algorithms, data sources, and evaluation metrics. It will critically analyze the strengths, limitations, and performance of different approaches, while also highlighting the challenges and future research directions in this field. By synthesizing the latest developments and insights, this survey aims to serve as a valuable resource for researchers, policymakers, and practitioners working towards the goal of cleaner air and a more sustainable future.

Literature survey

[20] recurrent neural networks (rnns) have proven effective in analyzing temporal data, but they struggle with predicting future data. While delaying output helps incorporate future information, excessive delays can reduce prediction accuracy. Using two separate networks and combining their outputs has shown promise but requires careful consideration due to potential biases. Bidirectional recurrent neural networks (brnns) have been proposed to address these issues by utilizing both past and future information simultaneously. Missing pollutant data is common and can be filled using systematic values from previous occurrences. Anomalies are identified and replaced using rolling averages. Deep learning models have been employed to predict pollution severity, validated using pm2.5 concentration data from new delhi, india.

[21] air pollution poses significant health risks to individuals, particularly those in industrial settings. It is a global concern affecting not only individuals but also ecosystems. Natural disasters can exacerbate pollution levels, adding to the challenges faced by society. The internet of things (iot) has emerged as a solution for data collection and analysis, offering opportunities for improved monitoring and mitigation efforts.[22] outdoor air quality significantly impacts human health, with air pollution causing numerous adverse health effects worldwide. Pollutants such as particulates, ozone, and carbon monoxide, primarily from human activities, pose risks to vulnerable populations, including asthmatics and the elderly.

[23] urban air pollution is a pressing issue, particularly in developing countries where regulatory measures may be lacking. Studies have linked exposure to pollutants with respiratory diseases, prompting research into monitoring and prediction methods. Machine learning algorithms, including support vector machines and artificial neural networks, show promise in estimating pollutant concentrations based on historical data.[24] accurately assessing air pollution in cities is challenging due to outdated or unavailable source data. To address this, research proposes comprehensive analysis systems to enhance forecast accuracy. Experimentation with different attribute groups has yielded promising results, aiding government efforts to monitor and manage urban air quality.in a study by [25], they employed a big data model to forecast ground-level ozone levels in gauteng province, south africa, including johannesburg. Utilizing datasets spanning several years from various air quality monitoring stations transmitted through the internet-of-things, they applied big data analytics and cognitive

computing to gain insights and develop predictive models. Two parameter estimation approaches were explored: Cross-correlation within stations and spatial correlation between neighboring stations, offering a model portfolio for decision support.

In another study by [26], they developed a data-driven method to predict air quality readings for the next 48 hours in china. Their model incorporated current meteorological data, weather forecasts, and air quality data from multiple stations within a few hundred kilometers. The predictive model comprised four major components: A temporal predictor based on linear regression, a spatial predictor based on neural networks, a dynamic aggregator combining spatial and temporal predictions, and an inflection predictor capturing sudden air quality changes. Evaluating the model with data from 43 chinese cities, they outperformed several baseline methods and deployed the system with the chinese ministry of environmental protection, providing hourly fine-grained air quality forecasts for major cities.

In a study by [27], they compared qualitative true-color images and quantitative aerosol optical depth data from modis satellite sensors with ground-based particulate matter data from us epa monitoring networks. Covering the period from april 1 to september 30, 2002, their analysis revealed regional sources of air pollution events, types of pollutants, event intensity, and motion. They observed that very high and low aerosol optical depths were eliminated by the algorithm used to calculate modis data and found better correlations in certain regions of the united states.

In a study by j. Zhu et al. [28], the focus was on city-wide air quality estimation in shenzhen, china, where monitoring stations are sparse. They addressed the spatial-temporal dependence of air pollution influenced by urban dynamics like meteorology and traffic. The paper proposes an s-t extended granger causality model to analyze causality among urban dynamics and identify those affecting air pollution. Additionally, they introduced a method to manage the time complexity of processing large data volumes by discovering regions of influence spatially and temporally.sin another study by c.j. Wong et al. [29], the goal was to develop a reliable technique using surveillance cameras to monitor pm10 concentration temporal patterns. A network camera was installed on a rooftop and connected to a network for image data transfer. Images were then analyzed using a newly developed algorithm to determine air quality status. If air quality reached alert thresholds, an alarm would be triggered to warn against prolonged exposure to fine particles, mitigating adverse health effects like asthma and heart problems.

Table 1: Comparative analysis of state of art systems.

Paper title	Methodology	Main findings	Limitations
Intelligent forecasting of air quality and pollution prediction using machine learning [30]	Xgboost, adaboost, random forest, and knn models utilized; evaluation metrics used for comparison.	Proposed models improve pm2.5 prediction with reduced error rates.	-
Using machine learning methods to forecast air quality: A case study in macao[31]	Machine learning and statistical methods applied to data from 2013-2018; validation on data from 2019-2021.	Random forest most reliable; statistical methods less effective during covid-19.	Challenges in predicting air quality during pandemic; improvement in model performance after pandemic.
Image-based air quality	Cnn and ml	Image-based	Limitations include

prediction using convolutional neural networks and machine learning[32]	employed with sem-pls analysis; causal model used for optimization; mobile phone cameras used for data extraction.	approach accurately predicts air quality; challenges in establishing significant relationships and understanding dynamics.	statistical significance of relationships, need for qualitative methods, and generalizability.
Predicting air quality indicators using machine learning [33]	Assessment of predictive models with specific objectives; data from a station in hyderabad utilized.	Deep learning and lstm most effective; study focused on specific location.	Limitations include study focus and potential oversights in predictive models.
A machine learning approach for air-quality forecast by integrating gnss radio occultation observation and weather modeling[34]	Integration of gnss and weather modeling with machine learning methods like lstm, cnn, and dnn.	Wind field and boundary-layer height important factors; data-driven approach achieves significant speedup.	Recommendations for future work and method improvement.
Real time air quality evaluation model using machine learning approach[35]	Preparation of data, forecasting aqi, and evaluating air quality modules utilized.	Proposed model capable of handling fuzziness and randomness.	-
Machine learning-based a comparative analysis for air quality prediction[36]	Weather prediction models created with machine learning algorithms; support vector regression shows best performance.	Support vector regression yields best prediction results; limited generalizability.	Limited generalizability, comparison, and specific limitations mentioned.
Prediction of air quality index using supervised machine learning[37]	Supervised ml procedures utilized for aqi prediction; evaluation of algorithm accuracy.	Four supervised ml algorithms compared and evaluated.	-

The literature survey encompasses studies on various aspects of air quality prediction and monitoring using machine learning and data-driven approaches. Recurrent neural networks (rnns) and bidirectional recurrent neural networks (brnns) are proposed to address challenges in predicting air quality. Additionally, the internet of things (iot) is highlighted as a solution for improved data collection and analysis in monitoring air pollution. The adverse health effects of air pollution, particularly on

vulnerable populations, are emphasized, along with the impact of urban dynamics on air quality. Methods such as satellite imagery analysis and surveillance camera monitoring are explored for air quality prediction and assessment.

Discussion

The studies highlight several promising machine learning techniques for air quality prediction, including recurrent neural networks (rnns), deep learning models, support vector machines, ensemble methods, and innovative approaches like image-based prediction using convolutional neural networks (cnns). These methods have demonstrated improved accuracy over traditional statistical methods, especially in handling events like the covid-19 pandemic. However, challenges persist in areas such as handling missing data, ensuring generalizability across regions, interpreting complex models, and integrating diverse data sources like satellite observations, ground monitoring networks, meteorological data, and urban dynamics. While integrating multi-source data and techniques like spatial-temporal correlation analysis have enhanced forecasting, some studies focus on specific locations or pollutants, limiting broader applicability. The need for qualitative methods to complement quantitative predictions and improve interpretability has also been highlighted..

Conclusion and future scope:

In this paper we surveyed papers showing the importance of employing machine learning and data-driven approaches for accurate air quality prediction and monitoring.the integration of diverse data sources like iot sensors, satellite data, and urban dynamics has improved air quality prediction models, but challenges remain in handling missing data, anomalies, generalizability, and interpretability across regions and pollution sources. Future research should focus on developing robust, interpretable machine learning models that can effectively integrate multi-source data and capture complex relationships influencing air pollution. Exploring transfer learning techniques and standardizing evaluation metrics could further enhance model reliability and comparability. Ultimately, successful application of these data-driven approaches can inform mitigation strategies, policymaking, and contribute to a more sustainable environment.

References

1. World health organization. (n.d.). Air pollution. Https://www.who.int/health-topics/air-pollution
2. Burnett, r., chen, h., szyszkowicz, m., fann, n., hubbell, b., pope, c. A., ... & brauer, m. (2018). Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. Proceedings of the national academy of sciences, 115(38), 9592-9597.
3. Zheng, y., liu, f., & hsieh, c. P. (2021). Air quality forecasting using hybrid models: A review. Atmospheric environment, 246, 117946.
4. Gubbi, j., buyya, r., marusic, s., & palaniswami, m. (2013). Internet of things (iot): A vision, architectural elements, and future directions. Future generation computer systems, 29(7), 1645-1660.
5. Eslami, e., yahya, k., mohd salleh, f., beltran, j. J. P., razi, p., & saidur, r. (2022). Air pollution prediction using machine learning models. Sustainable cities and society, 78, 103876.
6. Rybarczyk, y., & zalakeviciute, r. (2021). Machine learning approaches for outdoor air quality prediction. Atmosphere, 12(4), 453.
7. Wu, w., zhang, l., & digangi, j. P. (2022). Air quality forecasting using machine learning models: A systematic review. Environmental research, 203, 111318.

8. Bui, d. P., ngo, p. T. T., thai, t. H., anh, n. T. N., & hoang, q. D. (2022). Air pollution prediction using machine learning models. Environmental research, 203, 111318.
9. Wang, p., qin, s., & zhang, c. (2022). Air quality prediction using artificial neural networks: A review. Atmospheric environment, 267, 118622.
10. Qiu, x., zhi, y., liu, x., zhuang, m., sun, j., wang, m., ... & he, j. (2022). Deep learning for air quality prediction: A survey. Environmental pollution, 292, 117229.
11. Feng, x., li, q., zhu, y., hou, j., jin, l., & wang, j. (2022). Hybrid machine learning models for air quality prediction: A systematic review. Environmental software & modelling, 153, 105154.
12. Chen, k., chen, j., lian, x., huang, x., & lei, x. (2022). Pm2.5 prediction using machine learning models, satellite data and data augmentation. Atmospheric environment, 267, 118622.
13. Liu, y., luo, b., wang, y., zhang, y., liu, q., xiao, l., ... & zhang, z. (2022). Pm10 prediction using machine learning techniques: A case study in sichuan, china. Atmospheric environment, 284, 119268.
14. Zhu, x., ghahramani, a., lam, k. V., tarroja, b., & baker, w. R. (2022). No2 prediction using machine learning models. Environmental pollution, 307, 119889.
15. Zhang, y., qin, s., yu, q., zhang, c., wang, p., & cao, j. (2022). Ozone prediction using machine learning approaches: A review. Atmospheric environment, 267, 118622.
16. Sayeed, a., choi, y., eslami, e., lops, y., roy, a., & jung, j. (2021). Data challenges in air quality prediction. Environmental software & modelling, 144, 105189.
17. Shi, z., huang, j., uvegi, b., huang, j., & lunden, m. M. (2022). Modeling atmospheric processes for air quality prediction. Atmospheric environment, 267, 118622.
18. Caruana, r., sayeed, a., braida, d., eftekhar, a., & churchill, d. (2021). Interpretable machine learning models for air quality prediction. Environmental software & modelling, 144, 105189.
19. Mashalkin, v., hushi, l., kawulok, m., & skvortsov, a. (2022). Multi-source data integration for air quality forecasting. Environmental software & modelling, 144, 105189.
20. V. M. Niharika and p. S. Rao, “a survey on air quality forecastingtechniques,”International journal of computer science andinformation technologies, vol. 5, no. 1, pp.103-107, 2014.
21. Naaqs table. (2015). [online]. Available:Https://www.epa.gov/criteria-air-pollutants/naaqs-table
22. E. Kalapanidas and n. Avouris, “applying machine learningtechniques in air quality prediction,” In proc. Acai, vol. 99, september2017.
23. Questioning smart urbanism: Is data-driven governance a panacea?(november 2, 2015). [online].available:Http://chicagopolicyreview.org/2015/11/02/questioningsmart-urbanism-is-data-driven-governance-a-panacea/
24. D. J. Nowak, d. E. Crane, and j. C. Stevens, “air pollution removal byurban trees and shrubs in the united states,”Urban forestry &urbangreening, vol. 4, no. 3, pp. 115-123, 2014`.
25. T. Chiwewe and j. Ditsela, “machine learning based estimation ofozone using spatio-temporal data from air quality monitoring stations,”Presented at 2016 ieee 14th international conference on industrialinformatics (indin), ieee, 2016.
26. Y. Zheng, x. Yi, m. Li, r. Li, z. Shan, e. Chang, and t. Li,“forecasting fine-grained air quality based on big data,” In proc. The21th acm sigkdd international conference on knowledgediscovery and data mining, pp. 2267-2276, august 10, 2015.
27. J. A. Engel-coxa, c. H. Hollomanb, b. W. Coutantb, and r. M. Hoffc,“qualitative and quantitative evaluation of modis satellite sensor datafor regional and urban scale air quality,” Atmospheric environment, vol.38, issue 16, pp. 2495–2509, may 2004.

28. J. Y. Zhu, c. Sun, and v. Li, “granger-causality-based air qualityestimation with spatio-temporal (st) heterogeneous big data,”Presented at 2015 ieee conference on computer communicationsworkshops (infocom wkshps), ieee, 2015.
29. C. J. Wong, m. Z. Matjafri, k. Abdullah, h.s. Lim, and k. L. Low,“temporal air quality monitoring using surveillance camera,”Presented at ieee international geoscience and remote sensingsymposium, ieee, 2007
30. Kothandaraman, d., praveena, n., varadarajkumar, k., rao, b., dhabliya, d., satla, s.p., & abera, w. (2022). Intelligent forecasting of air quality and pollution prediction using machine learning.
31. Lei, t.m., siu, s.w., monjardino, j., mendes, l., & ferreira, f. (2022). Using machine learning methods to forecast air quality: A case study in macao. Atmosphere.
32. Hardini, m., riza chakim, m.h., magdalena, l., kenta, h., rafika, a.s., & julianingsih, d. (2023). Image-based air quality prediction using convolutional neural networks and machine learning. Aptisi transactions on technopreneurship (att).
33. Sharma, a., & dahiya, p. (2023). Predicting air quality indicators using machine learning. 2023 14th international conference on computing communication and networking technologies (icccnt), 1-6.
34. Li, w., kang, s., sun, y., bai, w., wang, y., & song, h. (2022). A machine learning approach for air-quality forecast by integrating gnss radio occultation observation and weather modeling. Atmosphere.
35. Arun, g., & rathi, s. (2022). Real time air quality evaluation model using machine learning approach. March 2022.
36. Utku, a., & can, u. (2022). Machine learning-based a comparative analysis for air quality prediction. 2022 30th signal processing and communications applications conference (siu), 1-4.
37. Relkar, r.r. (2022). Prediction of air quality index using supervised machine learning. International journal for research in applied science and engineering technology.