

https://dspace.org/

Machine learning for air quality prediction and data analysis Review on recent advancements challenges and outlooks

Manal Karmoude; Brenton Munhungewarwa; Isaiah Chiraira; Ryan McKenzie; Jude Kong ...

https://hdl.handle.net/10539/47423
WOS

Downloaded from a DSpace repository. https://dspace.org/

Science of the Total Environment 1002 (2025) 180593

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

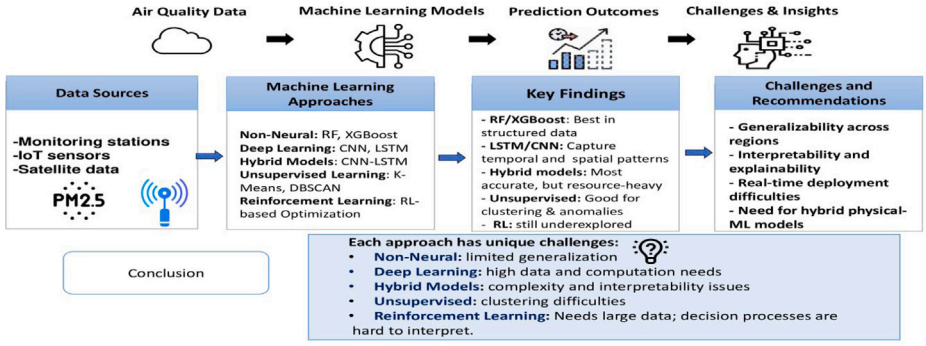
Review

Machine learning for air quality prediction and data analysis: Review on recent advancements, challenges, and outlooks

Manal Karmoude [a,b,*], Brenton Munhungewarwa [a], Isaiah Chiraira [c], Ryan Mckenzie [a], Jude Kong [d], Bevan Smith [e], Gelan Ayana [f], Nkosiphendule Njara [a], Thuso Mathaha [a], Mukesh Kumar [a], Bruce Mellado [a]

[a] School of Physics, University of the Witwatersrand, iThemba LABS, Johannesburg, South Africa
[b] Faculty of Sciences, Mohammed V University, Rabat, Morocco
[c] School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa
[d] Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
[e] School of Mechanical, Industrial and Aeronautical Engineering, University of the Witwatersrand, Johannesburg, South Africa
[f] School of Biomedical Engineering, Jimma Institute of Technology, Jimma University, Jimma, Ethiopia

GRAPHICAL ABSTRACT



HIGHLIGHTS

- Reviews 78 ML-based studies on air quality monitoring and prediction.
- Categorizes approaches into non-neural, deep learning, and hybrid models.
- Ensemble methods like RF and XGBoost perform well with structured datasets.
- Deep learning captures spatio-temporal patterns in pollution trends.
- Highlights challenges in generalizability, explainability, and real-time use.

ARTICLE INFO

Editor: P. Hopke

Keywords:
Air quality monitoring

ABSTRACT

Air quality is a critical determinant of human health, with severe consequences resulting from air pollution. The growing necessity for air quality monitoring has led to the adoption of IoT sensor networks, which provide real-time data for forecasting, issuing warnings, and informing public health interventions. In this context,

* Corresponding author at: School of Physics, University of the Witwatersrand, iThemba LABS, Johannesburg, South Africa.
E-mail addresses: manal.karmoude@cern.ch (M. Karmoude), brenton.munhungewarwa@cern.ch (B. Munhungewarwa), it.chiraira@cern.ch (I. Chiraira), ryan.peter.mckenzie@cern.ch (R. Mckenzie), jude.kong@utoronto.ca (J. Kong), bevan.smith@wits.ac.za (B. Smith), gelan.zewdie@utoronto.ca (G. Ayana), nkosiphendule.njara@cern.ch (N. Njara), thuso.mathaha@cern.ch (T. Mathaha), mukesh.kumar@cern.ch (M. Kumar), bruce.mellado.garcia@cern.ch (B. Mellado).

https://doi.org/10.1016/j.scitotenv.2025.180593
Received 30 April 2025; Received in revised form 21 September 2025; Accepted 21 September 2025
Available online 29 September 2025
0048-9697/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Supervised learning
Machine learning
Deep learning
Unsupervised learning
Forecasting air quality
Reinforcement learning

machine learning (ML) algorithms have proven to be powerful tools for enhancing air quality prediction and addressing monitoring challenges. However, a comprehensive review compiling the research space of ML for air quality is seldom available. This review analyzes over 70 recent studies that apply ML techniques to air quality monitoring, categorizing them based on the type of learning approach employed, with a focus on identifying the most effective algorithms in each category. The findings demonstrate that ensemble models such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost) consistently achieve high accuracy in structured datasets, while deep learning (DL) approaches like Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) excel in capturing temporal dependencies and spatial patterns in pollution forecasting. Unsupervised approaches like clustering and anomaly detection effectively enhance data quality and sensor calibration, whereas reinforcement learning shows promise in adaptive control scenarios, despite challenges related to computational intensity and interpretability. This review is highly significant, offering valuable insights for policymakers and researchers in developing strategies to mitigate air pollution and improve public health using advanced ML techniques.

Contents

1. Introduction ... 2
1.1. Background ... 2
1.2. Problem statement ... 3
1.3. Objectives ... 3
1.4. Contributions ... 3
2. Review scope and methodology ... 4
2.1. Search strategy ... 4
2.2. Screening process ... 4
2.3. Inclusion and exclusion criteria ... 4
2.3.1. Inclusion criteria ... 4
2.3.2. Exclusion criteria ... 4
2.4. Data extraction and synthesis ... 4
2.5. Study selection overview ... 4
3. Air quality prediction methods: From traditional models to ML approaches ... 4
4. Supervised ML in air quality ... 5
4.1. Tree-based algorithms for AQI prediction ... 5
4.2. Boosting algorithms for AQI prediction ... 5
4.3. Predicting pollutant concentrations using non-neural network algorithms ... 5
4.4. Calibration of low-cost air quality sensors using non-neural network ... 6
4.5. Deep learning algorithms for predicting pollutant concentration ... 6
4.6. Deep learning for AQI prediction ... 7
4.7. Hybrid approaches for particulate matter prediction ... 7
4.8. Hybrid approaches for AQI prediction ... 8
5. Comparative analysis of non-neural network, deep learning, and hybrid models ... 9
6. Common limitations of high-performing supervised ML models for air quality prediction ... 9
7. Unsupervised ML in air quality ... 10
7.1. Clustering techniques for air quality data analysis ... 10
7.2. Anomaly detection for sensor calibration and data cleaning ... 10
7.3. Challenges and limitations of unsupervised learning in air quality monitoring ... 11
8. Reinforcement learning in air quality ... 11
8.1. Challenges and limitations of reinforcement learning in air quality applications ... 11
9. Types of data used in air quality monitoring studies ... 11
10. Computational efficiency in air quality monitoring ... 11
11. Interpretability in air quality forecasting ... 12
12. Discussion ... 13
13. Conclusion ... 13
CRediT authorship contribution statement ... 16
Declaration of competing interest ... 16
Acknowledgments ... 16
Appendix ... 16
Data availability ... 16
References ... 16

1. Introduction

1.1. Background

As a result of technological advancements and rapid industrialization, air pollution has become a serious global problem. It is caused by a combination of gases and particles, such as $PM_{2.5}$, nitrogen oxides (NO_x), and sulfur dioxide (SO_2), which are released into the atmosphere and lead to major problems. The issue of air pollution is particularly severe in urban areas due to high traffic emissions, industrial activities, and energy consumption (Kunak, 2023). According to a study by the Centers for Disease Control, air quality in rural areas is often better than in urban areas. However, although pollutant levels are low, the study still detected measurable contaminants in rural areas (Oransi, 2023). Studies have shown that exceeding the World Health Organization (WHO) threshold limits for air pollutants can have severe consequences for human health (Ritchie and Roser, 2024). The WHO recommends maintaining annual mean concentrations of $PM_{2.5}$ below 5 $\mu g/m^3$, NO_2 below 10 $\mu g/m^3$, and the peak season

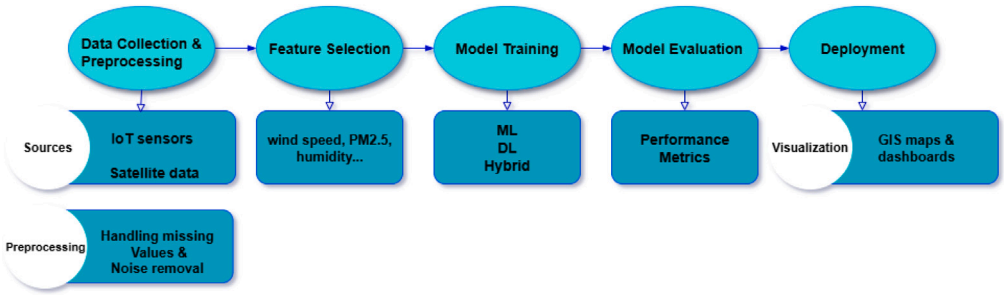


Fig. 1. Generalized ML workflow for air quality monitoring and forecasting.

8-hour mean ozone concentration below 60 μg/m^3 (World Health Organization, 2021b,a). According to the WHO, air pollution is the fourth leading cause of death worldwide, following cardiovascular disease, smoking, and poor diet (World Health Organization, 2021c). In 2019, air pollution, both ambient and household, was responsible for 6.7 million deaths globally, highlighting the significant health burden posed by poor air quality (World Health Organization, 2024). In addition to its impact on human health, air pollution also negatively affects both the environment and the economy (Pandey et al., 2020).

1.2. Problem statement

Understanding and having accurate information about air quality are critical for creating and evaluating strategies to reduce the harmful effects of pollution. However, real-time monitoring alone is not enough, accurate prediction of air quality is also important. Predicting air pollution trends allows for early warnings, enabling governments and individuals to take preventive actions before pollution levels become hazardous.

To address these concerns, smart city technologies such as the Internet of Things (IoT) and Information Communication Technology (ICT) have become increasingly important (Rakholia et al., 2023). IoT technologies play an important role in daily human life by introducing advanced solutions to various challenges. Today, IoT provides an innovative way to detect dangerous gases and monitor air quality through IoT-enabled sensors. IoT technologies not only enable real-time air quality monitoring through sensor networks but also facilitate predictive modeling by providing large-scale pollution data. These smart sensors continuously collect real-time information about air quality, such as pollutant levels, and immediately transmit this data to a central system or office over the Internet, helping to maintain a healthy and livable environment (Ali, 2022). Despite the importance of IoT in facilitating the collection of large amounts of air quality data, making sense of this data requires advanced analytical tools such as machine learning (ML).

ML has become increasingly popular in air quality prediction because of its high accuracy and ability to detect patterns, helping to provide early warnings for pollution-related health risks. When combined with IoT devices, ML enhances air quality monitoring by efficiently collecting and analyzing environmental data from various sensors. The generalized workflow for applying ML in air quality monitoring is illustrated in Fig. 1, showcasing key stages such as data collection, feature selection, model training, evaluation, and deployment.

Despite extensive research, a comprehensive literature compiling the knowledge space of ML for air quality is limited. Many studies focus on specific ML models or applications, but do not provide a clear and structured comparison of different ML approaches. A few existing reviews either categorize models based on learning paradigms or focus on model types, but only a handful of studies comprehensively analyze both perspectives.

1.3. Objectives

This review aims to bridge existing gaps in the literature by providing a structured and comprehensive analysis of ML approaches applied to air quality monitoring and forecasting. This includes the classification of ML techniques based on their learning paradigm: supervised, unsupervised, and reinforcement learning. Specifically, for supervised learning, we categorize models into non-neural network, deep learning (DL), and hybrid approaches, providing a detailed comparison of their strengths and weaknesses. By analyzing and synthesizing recent advancements in ML-based air quality monitoring, this review assesses the effectiveness of various ML techniques, highlighting high-performing models, key findings, and methodological advancements.

1.4. Contributions

This paper makes several key contributions to the field of ML-based air quality monitoring:

- **Comprehensive Categorization and Comparative Evaluation of Model Performance of ML Techniques:** We provide a structured classification of ML approaches used in air quality monitoring, organizing them into supervised learning (Section 4), unsupervised learning (Section 7), and reinforcement learning (Section 8) paradigms. Supervised learning is further divided into non-neural network models, deep learning models, and hybrid approaches. We analyze and compare the effectiveness of different ML models in predicting air quality. The results highlight that ensemble models such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost) consistently achieve high accuracy in structured datasets, while deep learning approaches like Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) excel in capturing temporal dependencies and spatial patterns in pollution forecasting. A comparative analysis between non-neural networks, deep learning, and hybrid models is provided in Section 5 to assess their strengths and weaknesses.
- **Analysis of Emerging Trends and Challenges:** The paper examines the latest advancements in ML-driven air quality prediction, including the increasing adoption of hybrid models that integrate multiple algorithms for enhanced accuracy. We also discuss key challenges such as computational complexity.
- **Overview of Data Sources and Features:** In Section 9, we provide a review of the types of data used in air quality studies, including pollutant concentrations, meteorological parameters.
- **Insights for Future Research:** By summarizing key findings from recent studies, this review provides actionable insights for researchers and policymakers, suggesting areas for further investigation. We identify promising directions such as optimizing ML models for real-time processing, integrating multimodal datasets for improved prediction accuracy.

This paper is organized as follows:

- Section 1 provides an introduction to air quality monitoring, highlighting the role of ML in improving prediction and analysis.
- Section 2 describes the scope and methodology of the review.
- Section 3 discusses traditional statistical models and compares them with ML-based approaches.
- Section 4 presents supervised ML techniques, categorizing them into non-neural network methods, deep learning models, and hybrid approaches, while evaluating their effectiveness in air quality monitoring.
- Section 5 compares the performance of non-neural network, deep learning, and hybrid models.
- Section 6 highlights common challenges and limitations of high-performing supervised ML models in air quality prediction.
- Section 7 explores unsupervised learning techniques, particularly clustering methods and anomaly detection.
- Section 8 examines reinforcement learning approaches applied to air quality monitoring and their limitations.
- Section 9 reviews the types of data used in air quality monitoring.
- Section 11 discusses interpretability in air quality forecasting.
- Section 10 discusses computational efficiency in air quality modeling.
- Sections 12 and 13 present the discussion and conclusion, respectively, synthesizing the findings, highlighting key insights, and emphasizing future research directions.

2. Review scope and methodology

2.1. Search strategy

To ensure a comprehensive review of relevant literature, a structured search strategy was employed. The objective was to identify studies that address key challenges in ML-based air quality monitoring, including forecasting pollutant concentrations, integrating heterogeneous data sources, and improving model scalability.

A systematic search was conducted across multiple academic and specialized databases, including ScienceDirect, Web of Science, Google Scholar, ResearchGate, IEEE Xplore, and SpringerLink. These databases were chosen due to their extensive coverage of computer science, artificial intelligence, and environmental science research.

To refine the search results, Boolean operators and keyword combinations were employed. The search terms included variations of "Air Quality Monitoring", "Air Pollution Forecasting", "Machine Learning", "Deep Learning", "Hybrid Models ", "Supervised Learning", "Unsupervised Learning", "Reinforcement Learning", "Spatio-temporal data", "Pollutant concentration", "Model optimization", "Real-time processing", and "Computational efficiency". These keyword combinations ensured a broad search scope, capturing a diverse range of studies relevant to ML applications in air quality monitoring.

2.2. Screening process

A multi-stage screening process was implemented to refine the selection of studies. Initially, all retrieved articles underwent a title and abstract screening to exclude irrelevant studies. Next, full-text reviews were conducted to assess the methodological rigor and relevance of each study. Duplicate articles retrieved from multiple databases were removed during this stage. Finally, only studies that met all inclusion criteria were selected for systematic analysis.

2.3. Inclusion and exclusion criteria

To maintain a high-quality review, specific inclusion and exclusion criteria were applied.

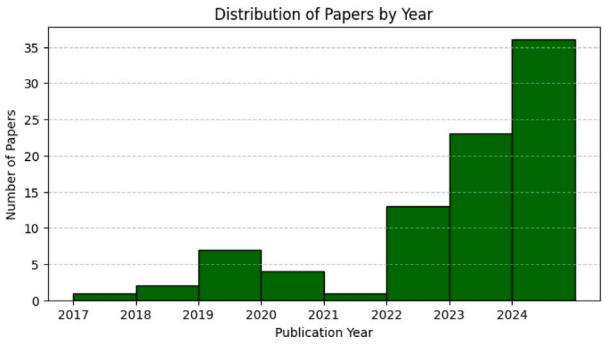


Fig. 2. Distribution of reviewed papers by year.

2.3.1. Inclusion criteria

Studies were included if they:

- Employed ML or deep learning techniques for air quality forecasting.
- Were peer-reviewed journal articles or reputable conference proceedings published between 2017 and 2024. Fig. 2 illustrates the publication years of the reviewed studies.

2.3.2. Exclusion criteria

Studies were excluded if they:

- Focused only on air quality monitoring without ML-based predictive modeling.
- Lacked real-world testing of ML models.
- Were non-peer-reviewed articles or preprints.
- Were duplicate studies retrieved from multiple databases.
- Addressed general environmental impact assessments without ML applications.
- Were written in languages other than English.

2.4. Data extraction and synthesis

For each included study, key data points were extracted, including publication year, ML models used, dataset characteristics, evaluation metrics, computational efficiency considerations, and key findings. The extracted data facilitated a comparative analysis of different models, allowing for an assessment of their strengths, weaknesses, and computational requirements. Studies were systematically categorized based on their methodologies and limitations. Performance metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) were examined to identify trends in model accuracy.

2.5. Study selection overview

The literature search initially identified over 150 relevant studies. Following the screening process and the application of inclusion and exclusion criteria, a final selection of 78 studies was made for systematic analysis.

3. Air quality prediction methods: From traditional models to ML approaches

Time series forecasting methods, such as AutoRegressive Integrated Moving Average (ARIMA), are widely used for air quality prediction. ARIMA is effective for modeling and forecasting short-term air quality trends when pollutant levels follow a linear pattern (Nguyen et al., 2024d). However, for long-term forecasting and capturing seasonal variations, Facebook Prophet, a powerful time series forecasting model,

has been shown to provide more interpretable and flexible predictions (Waqas et al., 2024). Traditional statistical methods often fail to account for nonlinear interactions between pollutants and environmental factors, limiting their effectiveness in accurately modeling air quality trends. Recent research has demonstrated that ML models such as RF outperform traditional methods by effectively incorporating non-linearity and interactions between meteorological features, improving the correction of meteorological influences on pollutant concentrations (Qiu et al., 2022).

ML techniques can analyze multivariate data, such as meteorological conditions and emission levels, to uncover complex patterns and relationships that traditional statistical models struggle to detect. ML, especially neural networks, is becoming increasingly popular for predicting air pollution levels. Artificial Neural Networks (ANNs) are particularly effective at capturing complex relationships between pollutants and factors like weather conditions and emissions (Nguyen et al., 2024a). These models have been successfully used to predict air quality in both the short and long term, making them a powerful tool for tackling air pollution challenges.

4. Supervised ML in air quality

Supervised learning has been applied in numerous studies (Vieru and Cărbureanu, 2024; Lyu et al., 2022) to address challenges related to air quality monitoring, whether for forecasting purposes, exploring the relationships between air quality and various pollution factors, examining spatiotemporal variations, or investigating how changes in certain factors influence air quality predictions and related concerns. Supervised learning includes a wide range of algorithms, from statistical models to more advanced ML techniques. In addition to traditional regression-based approaches, non-neural network algorithms have been widely employed due to their ability to handle structured environmental data efficiently. Among these, tree-based algorithms have demonstrated strong predictive capabilities, particularly for Air Quality Index (AQI) prediction.

4.1. Tree-based algorithms for AQI prediction

Many studies Mulomba and Choi (2024a), Bhalgat et al. (2019) have focused on predicting the AQI, a numerical index used to report air quality and measure air pollution levels. Tree-Based algorithms are commonly used and have shown superior performance. Among these, RF has consistently demonstrated the best performance, making it one of the most suitable algorithms for AQI prediction. RF is especially effective because it can handle non-linear patterns, making it great at capturing complex relationships between air quality variables. It is a flexible algorithm that can be used for a variety of tasks, including classification through the RF Classifier (RFC) and regression through the RF Regressor (RFR).

Bhalgat et al. (2019) found that RF outperformed Gaussian Naive Bayes and Linear Regression in AQI prediction. The study identified pollutant features such as PM2.5, PM10, and CO as key contributors to its success, confirming RF's reliability for historical AQI predictions. Similarly, RFR emerged as one of the top-performing algorithms in a study (Gupta et al., 2023) focused on AQI prediction for four major Indian cities (New Delhi, Bangalore, Kolkata, and Hyderabad). By utilizing Synthetic Minority Over-sampling Technique (SMOTE) to address dataset imbalance, RFR achieved a high accuracy rate and significantly lower RMSE scores compared to models like Support Vector Regression (SVR), further highlighting its effectiveness in handling complex data. Within the same country, another study (Oseni et al., 2024) investigated AQI variability across Delhi, Haryana, and Punjab, Gupta et al. demonstrated that RFR was the best-performing model, achieving near-perfect accuracy across Coefficient of Determination (R^2), Mean Squared Error (MSE), RMSE, and MAE. It effectively captured the complex relationships between air pollutants and meteorological factors, outperforming Category Boosting (CatBoost) and XGBoost, confirming its robustness in AQI prediction.

While RF has been shown to perform exceptionally well, Singh et al. (2024) demonstrated that the Decision Tree Regressor and K-Nearest Neighbors (KNN) outperformed RF and proved to be highly effective for fine-grained AQI prediction. The Decision Tree Regressor emerged as one of the best-performing models, demonstrating its ability to capture complex relationships between features such as CO, PM2.5, PM10, NO_2, and SO_2. Similarly, KNN also exhibited strong predictive performance. Although RF performed well, it was slightly outperformed by Decision Tree and KNN in the study. In contrast, Linear Regression and Logistic Regression showed lower predictive performance, further highlighting the advantages of tree-based and non-linear models for AQI forecasting.

4.2. Boosting algorithms for AQI prediction

Boosting algorithms are known for their ability to improve predictions by combining simple models, learning from mistakes, and making the overall performance much better. They have consistently demonstrated exceptional performance in AQI prediction tasks. Algorithms such as Adaptive Boosting (AdaBoost) and Gradient Boosting have been widely applied in various studies, often outperforming other traditional ML methods. Vieru and Cărbureanu (2024) evaluated multiple ML algorithms for AQI forecasting in Romanian cities, where AdaBoost and Gradient Boosting emerged as the top-performing models, delivering the most accurate and reliable predictions. Their ability to handle noisy datasets, capture complex relationships among features, and minimize prediction errors makes them some of the most reliable models for air quality forecasting.

Building on this success, more advanced boosting algorithms like XGBoost have continued to raise the bar, achieving even greater performance. In a study Kumar and Pande (2022) predicting AQI across 23 Indian cities, XGBoost achieved remarkable accuracy and low RMSE by leveraging pollutant data ($PM_{2.5}$, PM_{10}, CO, NO_2, SO_2) as inputs, significantly outperforming RF.

In a separate study Ravindiran et al. (2023), CatBoost, another robust boosting algorithm, was shown to deliver highly precise AQI predictions using both particulate matter and gaseous pollutants. CatBoost achieved the highest accuracy among all models. The study identified $PM_{2.5}$ and PM_{10} as the most significant contributors to AQI. CatBoost performed once again exceptionally well in regression tasks when predicting AQI status in multiple South African cities (Morapedi and Obagbuwa, 2023), achieving superior RMSE values. Meanwhile, RF and Decision Tree achieved perfect accuracy in AQI classification tasks, indicating the capability of boosting algorithms to predict AQI effectively even in diverse and complex environments.

4.3. Predicting pollutant concentrations using non-neural network algorithms

While non-neural network algorithms have proven effective in forecasting overall air quality through AQI prediction, these models are also widely employed for estimating the concentration of individual pollutants. Many studies Liu et al. (2023), Zhu et al. (2024) have used supervised algorithms to predict pollutant concentrations, with several demonstrating that RF consistently outperforms other models due to its ability to handle complex interactions and diverse datasets. For example, a study Lyu et al. (2022) conducted in the Beijing-Tianjin-Hebei (BTH) region of China utilized RF to investigate the spatiotemporal variations of air pollutants and predict ozone (O_3) levels. The study highlighted RF's superior performance over Decision Tree regression, demonstrating its ability to accurately capture relationships between meteorological factors, such as solar radiation and temperature, and ozone formation. This showcases RF's capability to effectively model complex environmental interactions. Similarly, RF demonstrated strong performance in predicting $PM_{2.5}$ concentrations using

spatiotemporal data from 10 cities across the Jing-Jin-Ji region of China (Ma et al., 2023). Another ensemble learning method, Extremely Randomized Trees (ExtraTrees), also performed well in this study, achieving results comparable to RF. These findings highlight the adaptability and effectiveness of ensemble models like RF and ExtraTrees in addressing region-specific air quality challenges, particularly during winter when pollution levels peak. While ensemble methods like RF and ExtraTrees have shown outstanding performance, simpler tree-based models have also demonstrated their effectiveness in air quality prediction. For instance, a study Kim (2023) conducted in Korean cities such as Yeosu, Gwangyang, and Suncheon employed Decision Tree Regression (DTR) for $PM_{2.5}$ prediction. The model outperformed Multiple Linear Regression (MLR) and SVR, demonstrating its strength in capturing non-linear relationships between pollutants and other influencing variables. These findings suggest DTR's suitability for real-time air quality management in regions with highly variable pollution levels.

Boosting algorithms have also demonstrated exceptional performance in predicting pollutant concentrations (Doreswamy et al., 2020). For instance, Gradient Boosting Machines (GBM) were applied in Christchurch, New Zealand, to forecast short-term $PM_{2.5}$ peaks (Miskell et al., 2019). The model proved highly effective in providing real-time predictions for short-term exposure mitigation, emphasizing the critical role of meteorological factors such as wind gusts and atmospheric pressure in shaping local air pollution dynamics.

Similarly, a study Morapedi and Obagbuwa (2023) conducted in South African cities found that CatBoost outperformed XGBoost in predicting $PM_{2.5}$ concentrations, achieving lower RMSE values. The study further demonstrated CatBoost's adaptability in leveraging data from neighboring cities to accurately estimate pollutant concentrations in regions lacking monitoring stations, highlighting its potential for enhancing air quality monitoring in areas with limited infrastructure. Further evidence of boosting algorithms' effectiveness comes from research in Ho Chi Minh City, Vietnam (Rakholia et al., 2024), where Light Gradient Boosting Machine (LightGBM) emerged as the top-performing model for short-term $PM_{2.5}$ forecasting. The study highlighted LightGBM's ability to efficiently handle complex datasets while incorporating meteorological, temporal, and rolling mean features. Its performance surpassed other models, including XGBoost, LSTM, and CNN-LSTM, demonstrating LightGBM's computational efficiency and accuracy in urban air quality management. Several studies Rovira et al. (2022), Munir et al. (2022) have also demonstrated the effectiveness of other traditional ML approaches in predicting pollutant concentrations. For instance, SVR was applied to estimate black carbon (BC) concentrations using various input features, including NO2, ultrafine particles, and PM2.5, in Barcelona, Spain Rovira et al. (2022). The model demonstrated its ability to provide reliable proxy estimates in urban environments where direct BC monitoring is limited. This study underscores SVR's utility in filling data gaps for non-regulated pollutants.

Ensemble models such as RF and ExtraTrees, along with boosting algorithms like GBM, CatBoost, and LightGBM, have demonstrated strong performance in air quality prediction. Additionally, tree-based methods like DTR have proven effective. Among these, LightGBM stands out as the most computationally efficient and accurate model for forecasting urban air quality.

4.4. Calibration of low-cost air quality sensors using non-neural network

Nowadays, low-cost sensors are becoming a popular way to monitor air quality because they allow measurements at more locations at a lower cost. These devices are small, lightweight, and typically cost less than ten percent of conventional high-precision sensors (Alvear-Puertas et al., 2022; Xayasouk et al., 2020). However, data from these sensors is not always reliable due to measurement errors and performance degradation over time, which can be affected by factors such as humidity or water droplets. ML-based calibration improves the accuracy of low-cost sensor measurements (Biagi et al., 2024), making them approach the reliability of reference-grade instruments. Many studies Wang et al. (2023), Apostolopoulos et al. (2023) have demonstrated the superiority of RF in calibrating low-cost sensors, showing consistent success. For example:

- In Ionascu et al. (2024), RF achieved strong calibration for PM_1, $PM_{2.5}$, and PM_{10} sensors, outperforming traditional methods like Multivariate Linear Regression (MLR), which struggled with non-linear relationships. However, RF faced challenges at low pollutant concentrations like for NO and SO_2, suggesting room for improvement in urban settings.
- In large networks, RF improved calibration across multi-hop systems (Vajs et al., 2023), boosting NO_2 correlation by 0.35 and cutting PM_{10} errors by approximately 20.56 $\mu g/m^3$.

While RF is reliable, newer models have surpassed it in some cases:

- Decision Trees outperformed RF and Support Vector Machine (SVM) in one study (Alvear-Puertas et al., 2022), achieving near-perfect alignment with reference data with low computational cost, ideal for portable IoT systems. Outlier removal like One-Class SVM further improved data quality.
- Boosting algorithms (Gradient Boosting, XGBoost) and SVR also excelled. These models boosted indoor $PM_{2.5}$ calibration (R^2 = 0.85–0.90), reduced errors, and fixed sensor underestimation. Performance varied by location (lunchrooms vs. classrooms), highlighting the need for environment-specific calibration (Chojer et al., 2022).

Fig. 3 shows the top supervised ML algorithms used in air quality monitoring. Boosting algorithms like XGBoost, AdaBoost, and CatBoost, along with RF, have consistently delivered the best predictive performance, often surpassing traditional models by effectively capturing complex relationships between air pollutants. SVM and KNN have also performed well in various studies Wu et al. (2017), Singh et al. (2024), especially when applied to structured air quality datasets. A summary of studies where non-neural network algorithms achieved better performance than other approaches in specific applications is provided in Tables A.4 and A.5 (Appendix).

4.5. Deep learning algorithms for predicting pollutant concentration

Neural networks have also been shown to be effective in solving air quality monitoring challenges due to their ability to process complex data and identify intricate patterns. Typically, when forecasting future air quality, time-series data is used. In this context, LSTM networks excel in analyzing sequential data (Cican et al., 2023), a critical requirement for air quality prediction.

An LSTM-based system forecasted indoor CO_2 concentrations over 120 min intervals, outperforming basic Recurrent Neural Networks (RNNs) (Rescio et al., 2023). The model integrated real-time environmental data (PM levels, humidity) and wearable sensor inputs (physical activity, heart rate), achieving a 50% accuracy boost when activity data was included. This highlights LSTMs' unique capacity to fuse temporal trends such as occupancy patterns with spatial inputs, enabling dynamic HVAC control and energy savings. LSTMs' dominance is evident in direct comparisons with other advanced architectures. Xayasouk et al. (2020) focused on Seoul's $PM_{2.5}$ and PM_{10} levels, demonstrated that LSTMs outperformed Deep Autoencoders (DAE), achieving lower RMSE values and faster processing times. While $PM_{2.5}$ predictions were slightly less accurate due to particulate variability, LSTMs effectively captured Seoul's complex pollution dynamics, integrating meteorological data (temperature, wind) to enhance robustness. In contrast, DAEs struggled with computational inefficiency, underscoring LSTMs' balance of accuracy and scalability.

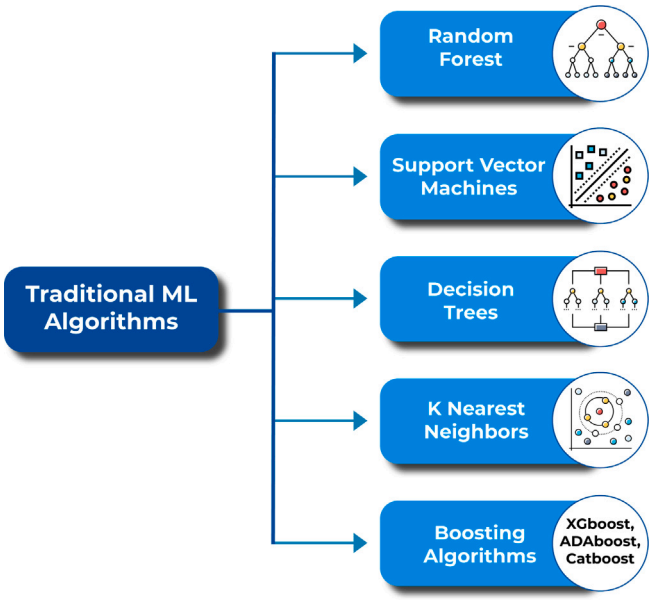


Fig. 3. Most commonly used and high-performing non-neural network algorithms in air quality prediction.

While LSTMs have demonstrated strong performance in air quality forecasting, other ML models have also shown competitive results. For instance, ensemble models combining ANN and Bidirectional Long Short-Term Memory (Bi-LSTM) achieved the highest accuracy in predicting PM_{10} levels in Sri Lanka (Mampitiya et al., 2023), significantly outperforming standalone LSTMs. These ensemble models exhibited exceptional forecasting capabilities, capturing both peak and trough variations more effectively than LSTMs.

Similarly, a multi-output Neural Basis Expansion Analysis for Time Series (N-BEATS) neural network was used for air quality forecasting in Ho Chi Minh City, Vietnam (Rakholia et al., 2023). This model leveraged deep stacked fully connected layers to predict NO_2, SO_2, CO, and O_3 concentrations simultaneously. Compared to existing models, N-BEATS achieved lower RMSE values and higher correlation coefficients (0.7–0.85), outperforming traditional forecasting techniques. The ability to predict multiple pollutants at once provided an efficient alternative to separate models for each pollutant, making it suitable for real-time urban air quality monitoring.

4.6. Deep learning for AQI prediction

In recent studies, LSTM networks have demonstrated robust performance in predicting the AQI, particularly due to their ability to model temporal dependencies in time-series data. For instance, LSTM outperformed traditional regression and classification models such as RF, SVR in forecasting pollutant levels, achieving high accuracy in capturing seasonal weather patterns and anthropogenic factors influencing AQI (Ugwu et al., 2024). The model effectively balanced computational efficiency and precision, making it suitable for real-time monitoring applications.

Similarly, Liu et al. (2024) reported that LSTM achieved the highest goodness-of-fit for regression-based AQI predictions in Jinan, outperforming LightGBM and Backpropagation Neural Networks (BPNN), especially when integrating meteorological variables like temperature and wind speed.

However, limitations persist. LSTMs require substantial computational power and careful tuning to prevent overfitting. Since they function as a "black box", their interpretability is limited compared to simpler models, posing a challenge for policy-making. Additionally, their performance can vary by location, as observed in Jinan (Liu et al., 2024), necessitating localized training data for optimal results.

CNN has also demonstrated exceptional performance due to its ability to capture spatial patterns and complex relationships in meteorological and air quality data. In a study Kalantari et al. (2024) conducted in Zabol, CNN achieved the highest accuracy, surpassing RF and other shallow learning methods. Another strong performer, however, was RF, which, despite being a shallow learning model, provided competitive results due to its robustness in handling diverse environmental features. While CNN excels in automatic feature extraction and classification, RF remains a reliable choice for AQI prediction, particularly when computational efficiency and interpretability are prioritized. Tables A.6 and A.7 (Appendix) present studies in which deep learning models outperformed other approaches. The most commonly used and high-performing deep learning algorithms for air quality prediction along with their typical applications are summarized in Fig. 4.

4.7. Hybrid approaches for particulate matter prediction

Hybrid models have gained increasing attention in air quality forecasting due to their ability to utilize the strengths of different computational approaches. By integrating multiple algorithms, these models enhance predictive accuracy, effectively capture complex spatio-temporal patterns, and improve robustness in handling time-series data. In particular, models combining CNN for spatial feature extraction and LSTM Networks for temporal sequence modeling have demonstrated significant advantages over non-neural network algorithms.

One such approach is the Convolutional-LSTM (CLSTM) model, as presented in Sharma et al. (2020), which integrates CNN with LSTM to predict Total Suspended Particulates (TSP) concentrations. By incorporating meteorological variables such as wind speed, temperature, and solar radiation, this model outperformed standalone ML methods, including RF, Multiple Linear Regression, and standard LSTM networks. Results showed higher accuracy, particularly in urban areas, reinforcing the reliability of hybrid deep learning frameworks for real-time air quality monitoring and risk assessment. However, the study also highlighted challenges such as computational complexity and limited generalizability when applied to finer time resolutions.

Expanding on this hybrid approach, Yang et al. (2024) introduced a spatio-temporal CNN-LSTM-SE model for $PM_{2.5}$ forecasting in the Beijing-Tianjin-Hebei region. This enhanced framework incorporated Squeeze-and-Excitation (SE) blocks, dynamically adjusting feature importance to optimize model performance. This model surpassed baseline approaches like BPNN, LSTM, and RF. While the model balanced accuracy and interpretability effectively, the study emphasized the need for high-quality data and acknowledged the computational intensity associated with the hybrid framework.

Beyond spatio-temporal deep learning models, researchers have explored data decomposition techniques to further refine air quality predictions. Empirical Mode Decomposition (EMD) and Variational Mode Decomposition (VMD) were applied to preprocess air pollution datasets, improving the performance of hybrid models like EMD-Gated Recurrent Unit (EMD-GRU) and CNN-LSTM (Shankar and Arasu, 2023). These models significantly outperformed standalone deep learning models, with EMD-GRU reducing RMSE and MAE compared to standard GRU architectures. The findings reinforce that integrating decomposition techniques with deep learning enhances the model's ability to capture temporal lags and reduce noise, leading to more reliable air quality forecasts. However, scalability remains a challenge due to the increased computational requirements of these hybrid techniques.

Freeman et al. (2018) employed an LSTM-RNN hybrid to predict 8-hour ozone levels in Kuwait, achieving a MAE of 0.235 ppb for 24-hour forecasts. By prioritizing meteorological features (wind speed, solar radiation) via decision trees, the model minimized computational costs while surpassing ARIMA (MAE: 23.574 ppb) and Feedforward Neural Networks (FFNNs). Its reliability for horizons up to 72 h underscores

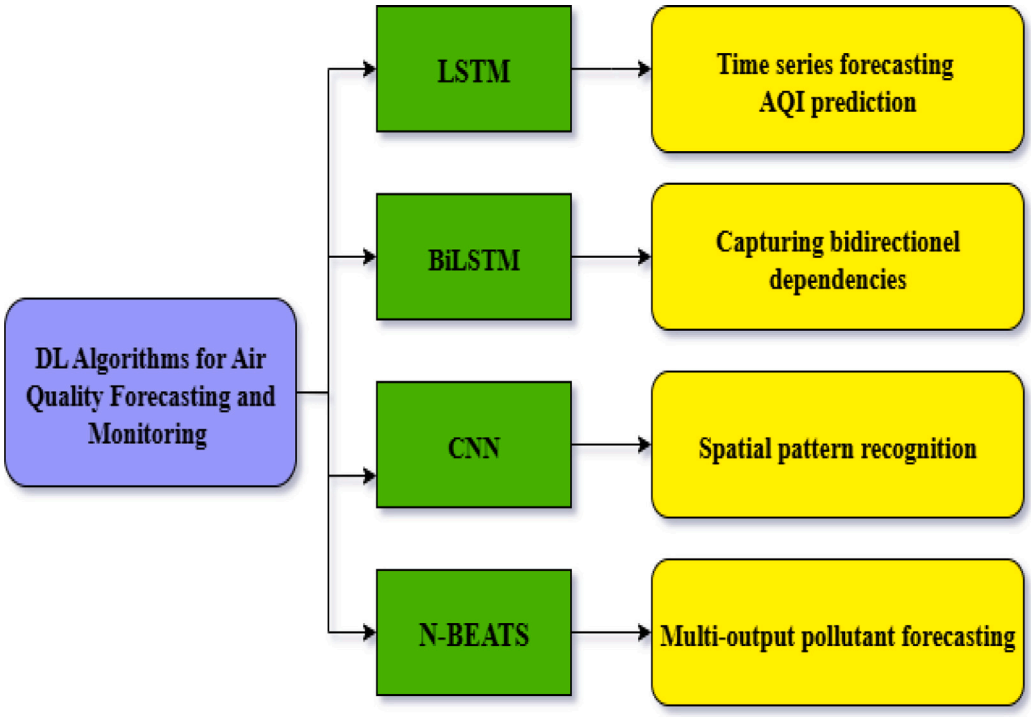


Fig. 4. Most commonly used and high-performing deep learning algorithms in air quality prediction.

LSTMs' suitability for real-time applications in regions with sparse monitoring infrastructure.

Integrating LSTM layers with a Deep Feedforward Neural Network (DFNN) for $PM_{2.5}$ prediction in Tehran demonstrated strong predictive performance (Karimian et al., 2019). Compared to traditional ML models like Multiple Additive Regression Trees (MART) and DFNN, the hybrid LSTM-DFNN model achieved the best performance, capturing 80% of the variability in $PM_{2.5}$ levels. The study highlighted the model's capability to effectively learn sequential patterns, particularly when incorporating meteorological variables alongside historical $PM_{2.5}$ data. However, a notable limitation was observed, as the model tended to underestimate high pollution levels and overestimate low pollution levels, indicating areas for further improvement in handling extreme values.

Furthermore, integrating clustering techniques with deep learning has proven effective in predicting extreme pollution events. Tamas et al. (2015) applied a Multilayer Perceptron (MLP) model enhanced with hierarchical clustering and a Self-Organizing Map (SOM) combined with k-means, allowing specialized MLP models to be trained for different clusters. While the baseline MLP model achieved strong overall accuracy, the hybrid clustering-based models (hMLP and kMLP) demonstrated superior performance.

4.8. Hybrid approaches for AQI prediction

Recent studies Zhao et al. (2023), Yasmin et al. (2023) have demonstrated that combining LSTM with other approaches significantly enhances air quality forecasting performance. In Ansari and Alam (2023), the BO-HyTS model integrates LSTM's ability to model nonlinear temporal dependencies with Seasonal Autoregressive Integrated Moving Average (SARIMA)'s strength in capturing linear seasonality and trends, while Bayesian Optimization fine-tunes the hybrid architecture. This combination effectively reduces prediction errors by leveraging both linear and nonlinear patterns in IoT-enabled pollution data, achieving the lowest error metrics compared to traditional models. Similarly, in Zhao et al. (2023), incorporating Spatial Autocorrelation (SAC) and Q-Learning-Based Bee Swarm Optimization (QBSO) with LSTM improved AQI predictions by reducing RMSE by up to 67.7%. Another study Yasmin et al. (2023), introduced AQIPred, a hybrid MLP-LSTM model that leveraged MLP's feature extraction with LSTM's temporal learning capabilities, achieving a high R^2. These findings reinforce that LSTM-based hybrid models effectively capture complex dependencies in air quality data, making them superior to standalone models in predictive accuracy and robustness.

BiLSTM-based hybrid models have further improved predictive accuracy by capturing bidirectional dependencies in time-series data. In Mahmoud and Mohammed (2024), the TCN-BiLSTM hybrid model, which integrates Temporal Convolutional Networks (TCN) with BiLSTM, effectively captured complex spatiotemporal dependencies, and outperforming both traditional and standalone deep learning models. Similarly, in Zhao et al. (2023), the BiLSTM model enhanced with SAC and QBSO achieved the best accuracy for Wuhan, demonstrating its effectiveness in leveraging spatial dependencies for air pollution prediction.

Some studies Kataria and Puri (2022), Nguyen et al. (2024d) have demonstrated that combining LSTM with multiple approaches significantly enhances predictive accuracy by effectively capturing both temporal dependencies and feature interactions. In Kataria and Puri (2022), the CNN-LSTM-BOA model, which integrates CNN for feature extraction, LSTM for temporal learning, and Bayesian Optimization (BOA) for hyperparameter tuning, achieved high accuracy across two datasets, outperforming baseline models. The study confirmed that Bayesian Optimization improves model stability and generalization, making the hybrid approach more robust.

Similarly, in Nguyen et al. (2024d), the ACNN-QPSO-LSTM-XGBoost model, incorporating Attention-based CNN, Quantum-inspired Particle Swarm Optimization (QPSO), and XGBoost, demonstrated strong predictive power in capturing irregular patterns in AQI data. This model effectively combined linear and non-linear modeling techniques, achieving high accuracy across pollutants, particularly for those with complex temporal dynamics like $PM_{2.5}$ and NO_2. Furthermore, the study highlighted that integrating optimization algorithms into deep learning frameworks enhances their ability to handle complex environmental data, making the approach robust and generalizable for urban air quality monitoring.

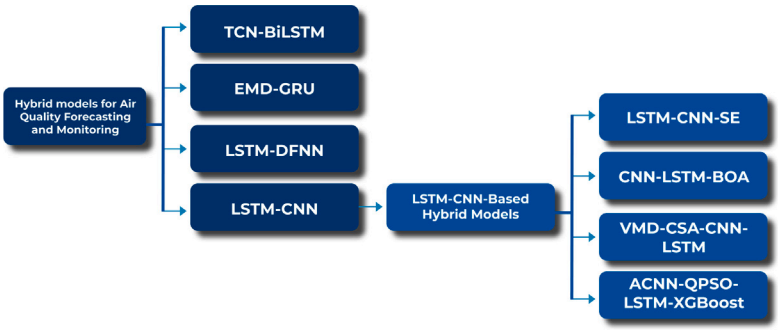


Fig. 5. Most commonly used and high-performing hybrid models in air quality prediction.

Additionally, in Guo et al. (2024), the VMD-CSA-CNN-LSTM model, which combines VMD for data decomposition, Chameleon Swarm Algorithm (CSA) for hyperparameter tuning, and CNN-LSTM for predictive modeling, achieved an adjusted R^2 and was externally validated across eight cities, confirming high performance. Moreover, the model effectively identified major pollutants specific to different regions, such as $PM_{2.5}$ in Beijing, PM_{10} in Lanzhou, and a combination of $PM_{2.5}$, PM_{10}, and O_3 in Shanghai and Wuhan, providing valuable insights for air quality management. These findings highlight that hybrid LSTM-based models, when combined with feature extraction, optimization, and decomposition techniques, not only improve overall accuracy but also enhance generalizability, robustness, and pollutant identification.

A summary of studies where hybrid approaches yielded better performance than other methods in various air quality prediction tasks can be found in Tables A.8 and A.9 (Appendix). The most commonly used and high-performing hybrid models for air quality prediction are summarized in Fig. 5.

5. Comparative analysis of non-neural network, deep learning, and hybrid models

The effectiveness of ML models in air quality prediction depends on their ability to handle complex environmental data, computational efficiency, and interpretability. Non-Neural Network, Deep Learning, and Hybrid Models each offer distinct advantages and trade-offs (Table 1).

Non-neural network models, such as RF and SVM, are computationally efficient and interpretable but rely heavily on manual feature selection. They perform well with structured datasets, which are data organized in a tabular format with clearly defined rows (instances) and columns (features), and small to medium-sized samples, but they struggle with capturing long-term dependencies and complex spatial–temporal relationships in air quality data. In contrast, DL models like CNNs, LSTMs and RNNs excel in automatically extracting features from large datasets and modeling intricate temporal variations in air pollution trends. Yet, they require significantly more computational power, are prone to overfitting, and suffer from a lack of interpretability, which limits their use in policy-making and regulatory applications.

Hybrid models, which integrate multiple approaches, aim to leverage the strengths of each. For example, CNN-LSTM models combine spatial feature extraction with sequential learning, improving predictive performance beyond what either approach achieves independently. Similarly, boosting algorithms such as XGBoost combined with neural networks enhance generalization and accuracy. However, hybrid models come at the cost of increased complexity and computational demands, making their deployment in real-time applications challenging.

In terms of performance, DL and hybrid models consistently achieve higher accuracy in air quality forecasting compared to non-neural network, particularly when dealing with large datasets and dynamic atmospheric conditions. But, non-neural network models remain relevant due to their faster training times, interpretability, and ease of implementation, making them suitable for scenarios where computational resources are limited or model explainability is crucial.

6. Common limitations of high-performing supervised ML models for air quality prediction

Despite major advancements in air quality prediction using ML, including traditional models, deep learning architectures, and hybrid approaches, several challenges remain. Even models that achieve high accuracy in research settings often face obstacles when it comes to broader applicability and real-world deployment.

One of the biggest issues is generalizability. Many studies Laukkarinen and Vinha (2024), Mulomba and Choi (2024a) develop and test their models using data from specific cities or regions. While these models perform well within those areas, they often struggle when applied to locations with different pollution sources, climate conditions, or emission patterns. This highlights the need for more diverse datasets that cover multiple geographic locations to ensure that ML models are robust across various environmental conditions.

Another key challenge is computational complexity. Deep learning models, such as LSTMs, CNN-LSTM hybrids, and decomposition-based approaches, require significant computing power for both training and inference (Mahmoud and Mohammed, 2024; Liu et al., 2024). Although they often outperform traditional ML models, their scalability remains a concern, especially for real-time applications or for use in regions where high-performance computing resources are limited. Similarly, ensemble methods like XGBoost and CatBoost, while effective, come with high computational costs, making them impractical for continuous real-time air quality monitoring.

Feature selection is also a limiting factor. Most models focus on standard air quality parameters like CO, NO, and meteorological variables such as temperature, humidity, and wind speed (Vm et al., 2020). An important aspect that is often overlooked is the role of additional predictors such as traffic density, industrial activity, population distribution, and land use. Incorporating these factors could improve the accuracy of air quality models, particularly in urban environments. In areas where human activities significantly influence air pollution levels, the exclusion of these factors can limit model accuracy. Some studies also lacked high-resolution meteorological data, making it harder for models to capture small-scale variations in air quality.

When it comes to real-time air quality monitoring, studies integrating IoT-enabled low-cost sensors faced difficulties with sensor reliability and calibration (Rescio et al., 2023). These sensors are prone to measurement drift and environmental interference, which can affect data accuracy. While some studies applied ML-based calibration techniques to correct these errors, results varied depending on the environment—indoor vs. outdoor settings, for instance, often showed inconsistent sensor performance.

Even high-performing ML models often suffer from overfitting and lack of interpretability. Deep learning and ensemble methods like CatBoost and XGBoost may provide highly accurate predictions, but they

Table 1
Comparison of ML approaches for air quality prediction.

Approach	Advantages	Challenges	Best use cases
Non-Neural Network	• Computationally efficient • Interpretable • Works well with structured datasets	• Struggles with long-term dependencies • Limited spatial–temporal analysis	• Ideal for resource-limited regions • Applications requiring transparency
Deep Learning	• Extracts features automatically • Captures complex temporal variations	• High computational cost • Prone to overfitting • Lacks interpretability	• Best for large datasets • Real-time air quality forecasting
Hybrid Approaches	• Combines strengths of multiple approaches • High predictive accuracy • Captures spatio-temporal patterns	• Computationally intensive • Complex deployment for real-time use	• Integrates multiple environmental factors • Generalizes across locations

tend to overfit when trained on small or imbalanced datasets. To address this, techniques like SMOTE are commonly used for dataset balancing (Gupta et al., 2023). However, their application can introduce additional computational complexity. Another major drawback is that these advanced models are often difficult to interpret, making it challenging for researchers, policymakers, and environmental agencies to understand how predictions are made. The lack of interpretability remains a significant barrier to integrating ML models into environmental policies.

Most studies Liu et al. (2023), Miskell et al. (2019) focus on short-term air quality forecasting (hourly or daily predictions), but long-term forecasting remains underexplored. The absence of models capable of capturing long-term patterns limits their effectiveness in sustainable air quality management strategies. Data availability and quality also pose challenges. Many studies Nguyen et al. (2024b), Singh et al. (2024) struggle with missing values in their datasets, often using imputation techniques to fill the gaps. However, these methods can impact model reliability.

7. Unsupervised ML in air quality

Unsupervised learning techniques, such as clustering and anomaly detection, have become valuable tools for improving the analysis, forecasting, and calibration of air quality monitoring systems.

7.1. Clustering techniques for air quality data analysis

SOM clustering, an unsupervised learning technique, was used to cluster air pollution data based on meteorological conditions rather than fixed time periods (Hulkkonen et al., 2022). This approach allows for the identification of distinct pollution patterns influenced by weather variables, rather than relying on arbitrary time-based groupings. By clustering data in this way, the study aimed to capture the non-linear relationships between pollutant concentrations and meteorological factors. Usually, researchers compare pollution levels based on the calendar, such as month-to-month or year-to-year comparisons. However, it was demonstrated that comparing pollution levels by calendar dates can give a misleading picture of air quality trends. In contrast, using SOM-based clustering provided a more accurate view by considering pollution changes under similar weather conditions.

K-Means clustering, another unsupervised clustering technique, was used to group similar air quality and meteorological data into clusters based on pollution levels and environmental factors (Ramirez-Alcocer et al., 2022). This approach helped identify patterns in COVID-19 case trends related to air pollution exposure by distinguishing indoor and outdoor pollution sources and ensuring a more structured dataset. By structuring the data before applying deep learning models, unsupervised clustering improved COVID-19 case prediction.

In addition to K-Means, Du and Siegel (2023) demonstrated that K-Means clustering was used to classify air pollution data into different phases, such as emission, plateau, and decay periods, while Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was employed to refine pollutant decay detection. Unlike K-Means, which requires a predefined number of clusters, DBSCAN is more effective in handling noise and short-term fluctuations in pollutant concentrations. However, DBSCAN has notable limitations when applied to heterogeneous air quality datasets. One of its main weaknesses is that it struggles with varying data densities, leading to misclassified outliers or fragmented clusters when pollution levels fluctuate across different regions. Additionally, DBSCAN's performance is highly sensitive to hyperparameter selection (Shetty et al., 2024).

To overcome these challenges, a study Nguyen et al. (2024b) introduced the Correlation-Based Girvan-Newman (CGN) algorithm, which outperformed DBSCAN in clustering air quality data. Unlike DBSCAN, which relies on density thresholds, CGN uses a graph-based approach that dynamically partitions sensors based on both spatial distance and correlation strength. CGN successfully grouped sensors into meaningful clusters, reducing classification errors and improving air quality monitoring compared to DBSCAN, which misclassified several sensors as noise.

While K-Means is a popular method for classifying air quality data, it has its limitations, one of the biggest being poor centroid initialization, which can lead to unstable and inaccurate clusters. To solve this, researchers integrated the Sine Cosine Algorithm (SCA) with K-Means, making centroid selection more precise and reliable (Vasudevan and Ekambaram, 2024). Unlike traditional K-Means, which randomly picks centroids, SCA-optimized K-Means carefully selects the best starting points, resulting in more stable and meaningful clusters. This improved clustering approach made it easier to categorize air pollution levels (Good, Moderate, and Poor) while significantly reducing classification errors.

In short, SOM cluster air pollution data based on meteorological conditions, improving trend analysis beyond time-based comparisons. K-Means and DBSCAN are widely used, with SCA-optimized K-Means enhancing cluster stability and CGN outperforming DBSCAN for sensor-based air quality classification.

7.2. Anomaly detection for sensor calibration and data cleaning

Air quality monitoring relies heavily on accurate sensor data, yet real-world conditions often introduce errors, fluctuations, and missing values. To improve data reliability, researchers have explored various unsupervised learning techniques to detect and correct anomalies before further processing. Park et al. (2023) demonstrated that combining multiple unsupervised learning methods improves the stability and accuracy of anomaly detection in air quality monitoring. Their study introduced a hybrid anomaly detection model that integrates LSTM Autoencoder (LSTM-AE) for detecting deviations based on data reconstruction, and One-Class SVM (OC-SVM) for refining outlier classification. Additionally, DBSCAN was applied as a preprocessing step to

group normal sensor readings before OC-SVM classification. Their findings emphasize the importance of pre-cleaning air quality data through both anomaly detection and data repair techniques, ensuring that sensor readings remain reliable despite fluctuations or sensor errors. By leveraging these methods together, their approach significantly reduced false positives and enhanced the robustness of anomaly detection models, making them more suitable for real-world applications.

While Park et al.'s approach focused on anomaly detection, another key challenge in air quality monitoring is that traditional sensor calibration models assume the input data is clean, which is rarely the case in practical scenarios. Zhivkov (2021) addressed this issue by incorporating anomaly detection as a preprocessing step before training ANNs for sensor calibration. The study found that removing the top 20 detected anomalies before training the model led to a significant improvement in calibration accuracy, increasing the R^2 score from 0.62 to 0.95 and reducing the MAE by 5.16%. These results highlight how unsupervised learning can enhance the performance of supervised models, particularly in improving sensor calibration by filtering out unreliable data. Expanding on these advancements, the AIrSense framework, introduced in an IoT-based air quality monitoring study (Rollo et al., 2023), took this approach a step further by integrating both anomaly detection and anomaly repair techniques.

Unlike Zhivkov's method, which removes anomalies, AIrSense identifies and corrects them using a Vector Autoregression (VAR) model, which reconstructs missing or corrupted data instead of simply discarding it. This prevents data loss, maintains data continuity, and ultimately improves sensor calibration.

7.3. Challenges and limitations of unsupervised learning in air quality monitoring

These studies highlight the crucial role of unsupervised learning in improving air quality monitoring. By combining anomaly detection, outlier filtering, and data repair techniques, researchers have significantly improved sensor accuracy and reliability. Despite their advantages, unsupervised learning methods in air quality analysis come with certain limitations. DBSCAN struggles with varying data densities and is highly sensitive to hyperparameters, leading to misclassified outliers. K-Means suffers from poor centroid initialization and requires a fixed number of clusters, reducing its adaptability. While graph-based methods like CGN improve clustering accuracy, they come with high computational costs. Some methods, such as One-Class SVM, rely on predefined anomalies, which may not always represent real-world variations. Additionally, removal-based approaches risk discarding valid extreme pollution events, leading to biased datasets. The computational complexity of hybrid models combining LSTM Autoencoder, OC-SVM, and DBSCAN makes real-time monitoring difficult.

8. Reinforcement learning in air quality

Unlike supervised and unsupervised learning, reinforcement learning has received less attention in air quality research. However, it has shown promising results in optimizing indoor air quality control and improving pollution forecasting. Reinforcement learning-based systems have helped reduce energy consumption while maintaining air quality, with Double Deep Q-Network (DDQN) performing better than traditional rule-based methods (Kim and Moon, 2023). In pollution forecasting, Q-learning has been used to fine-tune input selection and time delay in neural networks, leading to more accurate predictions and better handling of short-term fluctuations and noise in air pollution data (Chang et al., 2019).

8.1. Challenges and limitations of reinforcement learning in air quality applications

These findings highlight reinforcement learning's potential to enhance decision-making and improve predictive accuracy in air quality applications. However, reinforcement learning faces some limitations. It relies on large amounts of high-quality data (Chang et al., 2019), making it difficult to adapt to different environments without extensive retraining. Additionally, its decision-making process is often hard to interpret, making it unclear why certain actions are chosen.

9. Types of data used in air quality monitoring studies

Air quality monitoring data used in ML is mostly time-series data, recorded at regular intervals such as hourly (Ansari and Alam, 2023; Lyu et al., 2022), daily (Hasan et al., 2024), or even minutely. Some datasets also include aggregated or spatial data for broader analysis.

The key features in these datasets usually include pollutant concentrations such as $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, O_3, NO_x, NH_3, Benzene, Toluene, and Xylene, along with meteorological factors like temperature, humidity, wind speed, wind direction, and atmospheric pressure, which all play a role in how pollutants disperse (Rescio et al., 2023). Some studies Karimian et al. (2019) also consider solar radiation, cloud cover, and rainfall, while others Ravindiran et al. (2023) analyze heavy metals like lead, arsenic, cadmium, and nickel, as well as organic pollutants like benzo(a)pyrene for a more in-depth understanding of air quality.

In addition to pollutant and weather data, researchers often include geospatial and environmental factors, such as land-use patterns, population density (Kataria and Puri, 2022), road networks, and industrial activity, to improve the accuracy of air quality predictions.

Data for these studies typically comes from IoT-enabled sensors, government monitoring stations, and satellite-based remote sensing, creating structured datasets that can range from thousands to millions of records. Some models focus on spatiotemporal patterns (Lyu et al., 2022; Rautela and Goyal, 2024), while others use graph-based representations (Iskandaryan et al., 2023) to analyze relationships between different monitoring stations.

Before applying ML, researchers often preprocess the data to improve accuracy. This includes filling in missing values, normalizing measurements, engineering new features, and combining multiple data sources to enhance predictions. A typical air quality dataset consists of timestamps, pollutant levels, meteorological conditions, and AQI values, forming a structured format suitable for modeling and trend analysis.

The choice of ML model depends on the nature of the data. Time-series models like LSTM and BiLSTM are commonly used for forecasting air quality trends, while clustering and regression models like RF, Gradient Boosting, and DBSCAN help classify pollution levels, detect anomalies, and analyze trends.

Tables 2 and 3 summarize the data sources and features used in the reviewed studies. they also include the meteorological variables and data collection methods employed in each study.

10. Computational efficiency in air quality monitoring

To address computational constraints in air quality monitoring, Lightweight models and optimization techniques have been explored. An IoT-based AQ monitoring study in Ecuador (Alvear-Puertas et al., 2022) utilized Decision Tree classifiers instead of deep learning methods, significantly reducing computational load. Similarly, a Fog-Enabled AQ Prediction System (Pazhanivel et al., 2024) employed optimized Gated Recurrent Units for low-latency predictions on IoT devices, minimizing reliance on cloud computation. Some studies have also combined traditional time-series models with deep learning for efficiency, such as the BO-HyTS Model (Ansari and Alam, 2023),

Table 2
Part 1 – Summary of data sources and features in the reviewed studies.

Study	Data source/Location	Period	Pollutants measured	Meteorological features	Collection method
Ansari and Alam (2023)	CPCB, India	Jan 2015 – Jul 2020	PM2.5, PM10, CO, NOx, SO2, O3, VOCs, etc.	Not detailed	IoT sensors; hourly
Vieru and Cărbureanu (2024)	Romanian cities	1.5 years	SO2, NO2, O3, CO, PM10, PM2.5, Pb, etc.	Not specified	Historical AQ data
Lyu et al. (2022)	Beijing-Tianjin-Hebei, China	2014–2021	PM2.5, PM10, CO, SO2, NO2, O3	Temp, humidity, wind	TEOM, NASA MERRA-2
Miskell et al. (2019)	Christchurch, New Zealand	May 2016 – Oct 2017	PM2.5, PM10, CO, NO, NO2, SO2	Wind, temp, pressure	TEOM, nephelometers
Sridhar et al. (2022)	Indoor lab, India	Real-time	NH3, CO2, CO (MQ sensors)	Temp, humidity	Microcontrollers, WiFi
Shetty et al. (2024)	Urban regions (unspecified)	Multi-year	PM2.5, PM10	Wind, temp, direction	Ground stations
Sharma et al. (2020)	Queensland, Australia	Not specified	TSP (μg/m^3)	Wind, temp, RH, pressure, radiation	Ground stations
Gupta et al. (2023)	India (4 cities)	2015–2020	PM2.5, PM10, NOx, CO, VOCs	Not detailed	Monitoring stations
Rescio et al. (2023)	UPAI3 sensor + wearables	15 days	PM1, PM2.5, PM10, CO2	Temp, humidity, pressure	Smart t-shirt + indoor sensors
Alvear-Puertas et al. (2022)	Ibarra, Ecuador	2 months	CO, NOx, CO2	Temp, humidity, UV	Mobile IoT in vehicles
Xayasouk et al. (2020)	Seoul, South Korea	2015–2018	PM10, PM2.5	Rain, wind, sky, temp	25 stations
Cican et al. (2023)	Bucharest, Romania	2021–2022	NO2	Temp, humidity, wind	1 station (B-6)

Table 3
Part 2 – Summary of data sources and features in the reviewed studies.

Study	Data source/Location	Period	Pollutants measured	Meteorological features	Collection method
Rautela and Goyal (2024)	India (NASA MERRA-2)	2015–2022	Black carbon, SO4, dust, organics	Indirect	Satellite (reanalysis)
Nguyen et al. (2024b)	Sydney, Australia	Recent	PM2.5, PM10, others	Wind, humidity, rainfall	AQMS + PurpleAir
Kumar and Pande (2022)	India (23 cities)	2015–2020	PM2.5, PM10, CO, NO2, etc.	Not all included	AQ stations
Bhalgat et al. (2019)	Ahmedabad, India	2015	PM2.5, PM10, NOx, CO, etc.	Few	Kaggle repo
Mulomba and Choi (2024a)	Global (197 capitals)	Aug 2023–present	PMs, CO	12 meteo vars	World Weather Repo
Zhu et al. (2024)	Scotland	2016–2020	PM2.5, PM10, NO2	Temp, rainfall, land use	106 stations
Ravindiran et al. (2023)	Visakhapatnam, India	Recent	PMs, NOx, VOCs	10 meteo vars	Not specified
Doreswamy et al. (2020)	Taiwan	Daily/monthly	PM2.5	Meteorological	76 stations
Moursi et al. (2019)	Beijing, China	2010–2014	PM2.5	Dew point, temp, rain	UCI repo

which integrates SARIMA with LSTMs, dynamically optimizing weights to balance accuracy and computational cost. Additionally, a study leveraging CNN-LSTM-BOA (Kataria and Puri, 2022) demonstrated that Bayesian optimization effectively reduced hyperparameter tuning time while maintaining high predictive accuracy. To further enhance computational efficiency, feature reduction and data compression techniques have been employed. For instance, Principal Component Analysis and Autoencoders have been applied to lower AQ dataset dimensionality, thereby decreasing computational demands. for example, Shetty et al. (2024) used PCA to retain 75% of variance while reducing the number of input features, ensuring more efficient model training.

11. Interpretability in air quality forecasting

Most of the current machine learning models used in air quality research mainly focus on predictive performance rather than interpretability. As new methods are developed to improve predictive accuracy, transparent and interpretable models are often overlooked because the more complex models cannot be easily interpreted by humans (Ugwu et al., 2024).

To address this issue, several recent studies have integrated interpretability techniques into their forecasting frameworks. In the study by Yang et al. (2024), interpretability is achieved through a combination of feature selection and post-hoc explanation methods. A multi-stage feature selection process – comprising Pearson correlation, LASSO (Least Absolute Shrinkage and Selection Operator) regularization, and the Relief_F algorithm – is first applied to identify the five most relevant input variables influencing $PM_{2.5}$ concentrations. To interpret the model's predictions, the authors employ SHAP (SHapley Additive exPlanations), a model-agnostic method based on cooperative game theory, which quantifies the contribution of each selected feature to individual predictions. SHAP values are analyzed across different time periods and cities to provide both global and local interpretability, revealing how variables such as AQI (Air Quality Index), CO (Carbon Monoxide), and PM_{10} influence the forecast outcomes under varying seasonal and meteorological conditions. This approach enables the model to remain explainable despite its deep learning architecture and provides insights valuable for policy guidance in air pollution management.

In another study Jiménez-Navarro et al. (2024), the lack of transparency in deep learning models is further addressed by integrating both ante-hoc and post-hoc interpretability techniques. A novel Temporal Selection Layer (TSL) is embedded within LSTM and Feed-Forward networks to perform feature selection during training, highlighting the most relevant inputs while reducing model complexity. Additionally, SHAP is applied post-training to quantify the contribution of individual features, especially meteorological variables, on pollutant predictions. Together, these approaches enhance the explainability of the forecasting system.

Similarly, in the study Mulomba and Choi (2024a), interpretability was prioritized alongside predictive performance. To enhance trust and transparency in the model's predictions, three widely used interpretability techniques were employed: LIME (Local Interpretable Model-Agnostic Explanations), ELI5 (Explain Like I am Five), and PDPs (Partial Dependence Plots). LIME was applied to explain individual predictions for both regression and classification models, highlighting the positive or negative contribution of features such as $PM_{2.5}$, pressure, and humidity. ELI5 was used to compute feature importance weights, confirming that air pollutant features and the "feels like" temperature had the strongest influence on model outcomes. PDPs further illustrated

the global effect of each feature, revealing both linear and non-linear relationships with AQI categories. Together, these methods validated the model's behavior and ensured that predictions were interpretable and reliable for real-world application.

12. Discussion

Advanced ML significantly improves air quality monitoring, providing accurate and timely predictions. Ensemble methods like RF and XGBoost excel in structured datasets due to their robustness and handling of nonlinear relationships. Deep learning models, particularly LSTM and CNN, effectively capture temporal and spatial pollution patterns but require substantial computational resources.

Hybrid models combining deep learning and traditional algorithms leverage both spatial and temporal dynamics effectively, despite increased complexity and deployment challenges. Significant limitations identified include generalizability, interpretability, and computational efficiency. Addressing these issues through domain adaptation, lightweight architectures, and explainable AI are essential directions for future research.

Unsupervised learning offers valuable tools for anomaly detection and sensor calibration but faces challenges like hyperparameter sensitivity. These methods are not typically designed to predict specific pollution levels or AQI values as in supervised learning. However, they can complement supervised models by uncovering latent structures in the data, such as clusters of similar pollution conditions or relationships between pollutants and meteorological variables, which may improve feature engineering or pretraining strategies. Reinforcement learning shows potential for optimizing air quality management, though it demands extensive high-quality data and improved interpretability.

Overall, while ML techniques substantially enhance air quality predictions, addressing computational efficiency, model interpretability, and generalizability will be key to translating predictive insights into actionable public health and environmental strategies.

Despite these advancements, there are several limitations in the current body of literature that merit attention. First, issues related to fairness and bias remain largely unaddressed in the reviewed literature. None of the selected studies explicitly examined how ML models perform across diverse geographic or demographic contexts, such as urban versus rural settings or among underrepresented populations.

Second, while several studies utilized real-world data from low-cost sensor networks, particularly in urban and smart city environments, they generally lacked reporting on key deployment metrics, such as latency, system stability, or real-time inference performance. Real-time data refers to information that is continuously collected, processed, and made available with minimal delay. Additionally, the reviewed studies did not quantify computational trade-offs such as training time, memory usage, or model complexity.

While most reviewed studies did not explicitly address class imbalance, a few acknowledged its impact on model performance. Kumar and Pande (2022) applied SMOTE to mitigate the imbalance but reported that overfitting persisted in some models, suggesting that resampling techniques alone may be insufficient. Another study Lin et al. (2022) observed that the dataset was imbalanced for specific pollutants, particularly $PM_{2.5}$, which reduced prediction accuracy. For instance, when $PM_{2.5}$ was the target variable, the positive class represented only 16% of the training samples, limiting the model's generalizability. These findings emphasize the importance of properly handling imbalanced data in air quality modeling. Moreover, models trained on geographically localized datasets may suffer from regional overfitting, hindering their applicability to other areas. Future studies should prioritize cross-regional validation and adopt robust sampling strategies to enhance model reliability.

The reviewed studies span a variety of geographic contexts, with a strong concentration in Asia, particularly India and China. Other regions represented include Romania, Australia, South Korea, Malaysia, New Zealand, Ecuador, Scotland, Taiwan, and a few global datasets covering capital cities. However, regions such as the Middle East, North Africa, and much of Sub-Saharan Africa remain underrepresented. This geographic imbalance may limit the generalizability of findings to these areas. Future work should include more diverse settings to improve the global transferability and applicability of ML-based air quality prediction models.

We did not include a quantitative benchmarking comparison because the reviewed studies are based on diverse datasets collected from different geographical locations, with varying features and evaluation settings. As a result, directly comparing metrics like RMSE or MAE across these studies would not yield meaningful insights. Instead, we focused on a qualitative comparison to highlight which models performed best within their respective contexts.

Interpretability remains a cornerstone for real-world adoption of ML models in air quality forecasting. As discussed in Section 11, several recent studies have begun to incorporate both ante-hoc and post-hoc interpretability techniques such as SHAP, LIME, PDPs, and ELI5. These methods enable deeper understanding of feature influence and model behavior, which is crucial for policy decision-making and public trust.

Future research should focus on the following key areas:

- Enhancing model explainability: Developing transparent and interpretable ML models that provide insights into key contributing factors influencing air pollution.
- Improving computational efficiency: Implementing model compression techniques, federated learning, and edge computing to enable real-time processing with reduced computational costs.
- Integrating multimodal datasets: Combining air quality data with traffic patterns, industrial emissions, land use, meteorological parameters, and socioeconomic factors to improve prediction accuracy and contextual understanding.
- Developing adaptive ML models: Leveraging transfer learning and meta-learning to create models that can dynamically adapt to different environments with minimal retraining.
- Advancing real-time monitoring systems: Enhancing IoT sensor calibration, incorporating remote sensing data, and developing AI-driven anomaly detection systems for continuous air quality assessment.
- Exploring Transformer-based architectures: Investigating the application of Transformer-based models, such as the Temporal Fusion Transformer (TFT), Informer, and Autoformer, for capturing long-range temporal dependencies and improving the accuracy and interpretability of air quality forecasts.
- Implementing cross-regional validation frameworks to improve model generalizability across diverse geographical areas, as models trained on localized datasets may not perform well when applied in different regions with varying pollution sources.
- An important direction for future research involves strengthening the integration of ML-based predictions with actionable health and environmental interventions. Developing frameworks to directly link predictive outcomes to practical decision-making processes will enhance the real-world impact of ML-driven air quality solutions.

13. Conclusion

Air quality monitoring has become an essential area of research due to the increasing impact of pollution on human health and the environment. Rising industrialization, urbanization, and vehicle emissions have exacerbated air pollution, making it a critical challenge for governments and researchers worldwide. Traditional air quality assessment methods rely on physical monitoring stations, which, although accurate, are often expensive to maintain, have limited spatial coverage, and are unable to provide real-time predictions. To address these limitations, ML techniques have emerged as powerful tools for air quality

Table A.4
Non-neural network models leading air quality forecasting: A review of outperforming techniques (Part 1).

References	Objective	Dataset	Outperformed algorithm	Key findings
Goh et al. (2021)	Monitor vehicle cabin air quality, predict future conditions, and compare ML models for air quality prediction.	Data collected from Nissan Grand Livina and Toyota Vios, over 4800 km, 4 s intervals.	Support Vector Regression (SVR) (R^2: 0.9890–0.9981, RMSE: 2.5410, MAE: 0.4101).	SVR was most accurate for real-time prediction. Recirculation mode reduced pollutants but increased CO2 levels.
Dhope et al. (2024)	Forecasting System (RTAQSFS) in Pune and develop a real-time air quality forecasting system using IoT and ML models.	Hadapsar, Pune; 31 days of continuous monitoring, 2,232 readings.	Random Forest (RF) (R^2: 99.9%, lowest RMSE).	RF provided the most accurate predictions; strong pollutant-meteorology correlations.
Wang et al. (2023)	Develop and test calibration algorithms for low-cost air sensors in stationary and mobile settings.	New York City (stationary: Aug–Sep 2021, 1.6M points); Boston (mobile: Feb–Apr 2022, 130K points).	RF and Generalized Additive Model (GAM) (PM2.5 calibration, R^2 > 0.8); Linear models (NO2 calibration, R^2 0.7).	Stationary-trained models had poor transferability to mobile data. Regular retraining is required.

Table A.5
Non-neural network models leading air quality forecasting: A review of outperforming techniques (Part 2).

References	Objective	Dataset	Outperformed algorithm	Key findings
Vajs et al. (2023)	Develop a scalable, cost-efficient calibration propagation method for hybrid air sensor networks.	Novi Sad, Serbia; 10 low-cost sensors + 1 reference station; 1-min data over several months.	RF (NO2: r improved by 0.35, RMSE reduced by 6.82 μg/m^3; PM10: r improved by 0.14, RMSE reduced by 20.56 μg/m^3).	Improved accuracy across sensor networks; multi-hop calibration feasible at scale.
Bertrand et al. (2023)	Enhance Copernicus Air Quality forecasts using ML and MOS approaches.	CAMS data (2017–2019), pollutant concentrations from 1,535 stations.	RF (best RMSE reduction up to 48.1%), Gradient Boosting Machine (GBM) (best exceedance detection).	ML improved CAMS forecasts; Global models best for RMSE, Local models best for NO2.
Mulomba and Choi (2024a)	Develop a prediction framework using ML, focused on resource-limited settings.	World Weather Repository covering 197 capital cities worldwide.	RF (R^2 = 0.91, MSE = 0.0067) best for regression and classification.	Classification models more robust; explainable AI improved interpretability.

Table A.6
Deep learning-based air quality prediction studies: A review of outperformed models (Part 1).

References	Objective	Dataset	Outperformed algorithm	Key findings
Rautela and Goyal (2024)	Predict PM2.5 concentrations using deep learning techniques and analyze AI effectiveness in air quality monitoring.	NASA MERRA-2 Reanalysis Dataset (2015–2022), spatiotemporal variables (BCSMASS, DUSMASS25, OCSMASS, etc.).	CA achieved SSIM: 0.50–0.70, PSNR: 30 dB, MSE: 8–11 μg/m^3.	Strong predictive capabilities; accurately captured high PM2.5 levels in the Indo-Gangetic Plain but overestimated in northwestern regions.
Pazhanivel et al. (2024)	Develop a Fog-enabled real-time air quality monitoring and prediction system.	Multi-layer architecture: pollutant and meteorological data, real-time processing via fog nodes, cloud storage.	Seq2Seq GRU outperformed baseline models in multi-step air quality forecasting.	Real-time early warnings improved decision-making; efficient model deployment on Smart Fog Environmental Gateway.
Rakholia et al. (2023)	Forecast NO_2, SO_2, CO, and O_3 concentrations using a multi-output ML model.	Ho Chi Minh City, Vietnam (2021–2022); HealthyAir network; meteorological data from DarkSky API.	N-BEATS achieved correlation coefficients of 0.7–0.85 for most pollutants.	Efficiently predicted multiple pollutants with strong accuracy; integrated into the HealthyAir mobile app for real-time monitoring.

forecasting, enabling data-driven insights and predictive modeling that can enhance environmental decision-making. This review has examined various ML approaches used for air quality prediction, highlighting their effectiveness, challenges, and areas for improvement. The analysis covered non-neural network models, deep learning architectures, and hybrid techniques, showing that while deep learning models excel in capturing complex spatiotemporal patterns, traditional models remain relevant for their interpretability and efficiency.

The findings suggest that ensemble models such as RF and XGBoost continue to perform well in structured datasets, while deep learning approaches like LSTMs networks and CNNs demonstrate superior performance in sequential data analysis and pollutant concentration forecasting. Additionally, hybrid models integrating multiple techniques show promising advancements, offering enhanced predictive accuracy by leveraging spatial and temporal dependencies.

Despite these advancements, significant challenges remain. One of the major limitations is the generalizability of models across different geographic locations. Many ML models are trained on specific datasets and may not perform well when applied to new environments due to variations in climate, topography, and pollution sources. Addressing this challenge requires the incorporation of transfer learning and domain adaptation techniques to improve model robustness across diverse regions. Furthermore, the computational complexity of deep learning models poses a significant challenge, particularly for real-time air quality monitoring applications. Training deep learning models requires high-performance computing resources, which may not always be accessible, especially in low-resource settings. Developing

Table A.7
Deep learning-based air quality prediction studies: A review of outperformed models (Part 2).

References	Objective	Dataset	Outperformed algorithm	Key findings
Jiménez-Navarro et al. (2024)	Enhance interpretability and accuracy of air pollution forecasting using deep learning.	Graz, Austria (2014–2022); urban monitoring stations and ERA5-Land dataset.	TFF model performed best (MAE: 10.28 vs. 11.34 for FF; RMSE: 16.74 vs. 18.05 for FF).	TSL-enhanced models improved feature selection; wind speed and solar radiation were key predictors.
Kalantari et al. (2024)	Compare SL and DL methods for AQI prediction in Zabol, Iran.	Zabol City, Iran (2013–2022); daily PM10 and meteorological data.	CNN achieved highest accuracy (0.60); DL models outperformed SL models in AQI classification.	DL models better handled complex AQI classifications, but struggled with severe AQI categories.

Table A.8
Hybrid models leading air quality forecasting: A review of outperforming techniques (Part 1).

References	Objective	Dataset	Outperformed algorithm	Key findings
Sridhar et al. (2022)	Develop a cost-effective IoT-based platform for real-time air quality monitoring using hybrid ML techniques.	Indoor air quality data from MQ135 and MQ7 sensors (NH3, CO2, Benzene, CO) with meteorological parameters.	NARX/LSTM (best RMSE, outperformed baseline models like APNet).	Real-time predictions with strong accuracy; future work involves outdoor deployments.
Ly et al. (2019)	Develop an AI-based model for predicting NO_2 and CO concentrations from multisensor and weather data.	6,941 records (filtered from 9,357); pollutants (CO, NOx, NO_2, O3, NMHC) with meteorological data.	PSO-Optimized ANFIS (best predictions for NO_2 and CO).	PSO performed better than Simulated Annealing (SA); NO_2 predictions were most influenced by NMHC and NO_2-specific sensors.
Wei et al. (2024)	Integrate deep learning and Google traffic data to predict traffic-related air pollutants in urban environments.	Hong Kong traffic and air quality sensor data (April–July 2017); 50 m spatial resolution.	Deep Forest (highest R^2: NO = 0.72, NO_2 = 0.69).	Traffic data improved model accuracy by 5%–10%; fine spatial resolution enhanced localized pollution predictions.

Table A.9
Hybrid models leading air quality forecasting: A review of outperforming techniques (Part 2).

References	Objective	Dataset	Outperformed algorithm	Key findings
Lin et al. (2022)	Use deep learning and web search trends to predict air pollution levels in major U.S. cities.	EPA's Air Quality System and AirNow (2007–2018); Google Trends data for 152 pollution-related queries.	DL-LSTM (best accuracy: 87.6% for O_3, 83.4% for NO_2, 89.3% for $PM_{2.5}$).	Web search data enhanced prediction accuracy by up to 5% in some cities; search terms correlated with pollution spikes.
Natarajan et al. (2024)	Develop an optimized ML model for AQI prediction using feature selection and regression.	Public air quality data (2015–2020) from Kaggle; 26 major cities in India.	Grey Wolf Optimizer Decision Tree (GWO-DT) (avg. accuracy: 94.25% vs. 90.34% (SVR), 90.51% (KNN), 92.75% (RF)).	GWO-DT achieved highest AQI prediction accuracy; scalable to other regions.
Li et al. (2023)	Enhance fine-scale air quality assessments by integrating physical principles with deep learning.	Mainland China (2015–2018); 1,604 monitoring stations; pollutants (CO, NO_2, $PM_{2.5}$, PM_{10}, SO_2, O_3).	Deep Graph Network ($PM_{2.5}$ R^2: 0.87, PM_{10} R^2: 0.85, NO_2 R^2: 0.72, O_3 R^2: 0.78).	Outperformed MERRA2-GMI and WRF-Chem; strong generalization and spatial modeling.
Wang et al. (2024)	Develop a hybrid deep learning model for $PM_{2.5}$ prediction using spatial and temporal dependencies.	Beijing (2017–2020); 30 monitoring stations; pollutants ($PM_{2.5}$, PM_{10}, CO, O_3, NO_2) and meteorological factors.	VMD-GAT-BiLSTM (best short- and long-term forecasting accuracy).	Most accurate 1–6 h and 42 h forecasting model; effective spatiotemporal learning.

lightweight and optimized ML architectures that maintain predictive accuracy while reducing computational overhead is crucial for broader adoption. Another critical concern is the interpretability of deep learning models. While deep learning techniques achieve high accuracy in air quality forecasting, they often operate as "black-box" models, making it difficult to understand how predictions are made. The lack of transparency in these models limits their applicability in regulatory and policy-making frameworks, where decision-makers require clear justifications for recommended actions. To address this, future research should focus on explainable AI (XAI) techniques, such as SHAP (Shapley Additive Explanations) and attention mechanisms, to enhance the interpretability of ML-driven air quality forecasting systems. Moreover, real-time air quality monitoring using low-cost IoT sensors presents another challenge. While IoT-based air quality monitoring systems offer cost-effective and scalable solutions, they are prone to sensor drift, calibration errors, and environmental noise that can affect data reliability. Developing advanced sensor calibration techniques, leveraging ML for automatic correction of sensor biases, and integrating multi-sensor fusion approaches can significantly improve the accuracy of real-time air quality predictions.

As technology continues to evolve, ML will play a crucial role in enabling early warnings, policy interventions, and sustainable urban planning to mitigate the adverse effects of air pollution. By addressing the challenges of generalizability, computational efficiency, and interpretability, ML-driven air quality forecasting systems can significantly contribute to data-driven environmental governance, helping authorities implement timely pollution control measures and ultimately improving public health worldwide.

CRediT authorship contribution statement

Manal Karmoude: Writing – original draft, Methodology, Investigation, Conceptualization. **Brenton Munhungewarwa:** Validation. **Isaiah Chiraira:** Validation. **Ryan Mckenzie:** Validation. **Jude Kong:** Validation. **Bevan Smith:** Writing – review & editing. **Gelan Ayana:** Writing – review & editing. **Nkosiphendule Njara:** Validation. **Thuso Mathaha:** Validation. **Mukesh Kumar:** Validation. **Bruce Mellado:** Writing – review & editing, Supervision.

Funding

This work was supported by the International Development Research Centre (IDRC), the National Research Foundation (NRF) of South Africa, and the University of the Witwatersrand.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge the collaborative efforts of the South African Consortium for Air Quality Monitoring (SACAQM) team and the support provided by the School of Physics, University of the Witwatersrand.

Appendix

See Tables A.4–A.9.

Data availability

No new data were created or analyzed in this study. Data sharing is not applicable to this article.

References

Ali, A., 2022. A framework for air pollution monitoring in smart cities by using IoT and smart sensors. Informatica 46 (5), http://dx.doi.org/10.31449/inf.v46i5.4003.

Alvear-Puertas, V.E., Burbano-Prado, Y.A., Rosero-Montalvo, P.D., Tözün, P., Marcillo, F., Hernandez, W., 2022. Smart and portable air-quality monitoring IoT low-cost devices in Ibarra City, Ecuador. Sensors 22 (18), 7015. http://dx.doi.org/10.3390/s22187015.

Ansari, M., Alam, M., 2023. An intelligent IoT-cloud-based air pollution forecasting model using univariate time-series analysis. Arab. J. Sci. Eng. 49 (3), 3135–3162. http://dx.doi.org/10.1007/s13369-023-07876-9.

Apostolopoulos, I.D., Fouskas, G., Pandis, S.N., 2023. Field calibration of a low-cost air quality monitoring device in an urban background site using machine learning models. Atmosphere 14 (2), 368. http://dx.doi.org/10.3390/atmos14020368.

Bertrand, J.-M., Meleux, F., Ung, A., Descombes, G., Colette, A., 2023. Technical note: Improving the European air quality forecast of the copernicus atmosphere monitoring service using machine learning techniques. Atmos. Chem. Phys. 23 (9), 5317–5333. http://dx.doi.org/10.5194/acp-23-5317-2023.

Bhalgat, P., Pitale, S., Bhoite, S., 2019. Air quality prediction using machine learning algorithms. Int. J. Comput. Appl. Technol. Res. 8 (9), 367–370. http://dx.doi.org/10.7753/ijcatr0809.1006.

Biagi, R., Ferrari, M., Venturi, S., Sacco, M., Montegrossi, G., Tassi, F., 2024. Development and machine learning-based calibration of low-cost multiparametric stations for the measurement of CO2 and CH4 in air. Heliyon 10 (9), e29772. http://dx.doi.org/10.1016/j.heliyon.2024.e29772.

Chang, S.-W., Chang, C.-L., Li, L.-T., Liao, S.-W., 2019. Reinforcement learning for improving the accuracy of PM2.5 pollution forecast under the neural network framework. IEEE Access 8, 9864–9874. http://dx.doi.org/10.1109/access.2019.2932413.

Chojer, H., Branco, P., Martins, F., Alvim-Ferraz, M., Sousa, S., 2022. Can data reliability of low-cost sensor devices for indoor air particulate matter monitoring be improved? – An approach using machine learning. Atmos. Environ. 286, 119251. http://dx.doi.org/10.1016/j.atmosenv.2022.119251.

Cican, G., Buturache, A.-N., Mirea, R., 2023. Applying machine learning techniques in air quality prediction—A bucharest city case study. Sustainability 15 (11), 8445. http://dx.doi.org/10.3390/su15118445.

Dhope, T.S., Shaikh, A., Simunic, D., Patil, P.P., Wagh, K.S., Wagh, S.K., 2024. Real time air quality surveillance & forecasting system (RTAQSFS) in Pune city using machine learning-based predictive model. Proc. Eng. Sci. 6 (2), 505–512. http://dx.doi.org/10.24874/pes06.02.008.

Doreswamy, N., S, H.K., Km, Y., Gad, I., 2020. Forecasting air pollution particulate matter (PM2.5) using machine learning regression models. Procedia Comput. Sci. 171, 2057–2066. http://dx.doi.org/10.1016/j.procs.2020.04.221.

Du, B., Siegel, J.A., 2023. Estimating indoor pollutant loss using mass balances and unsupervised clustering to recognize decays. Environ. Sci. Technol. 57 (27), 10030–10038. http://dx.doi.org/10.1021/acs.est.3c00756.

Freeman, B.S., Taylor, G., Gharabaghi, B., Thé, J., 2018. Forecasting air quality time series using deep learning. J. Air Waste Manage. Assoc. 68 (8), 866–886. http://dx.doi.org/10.1080/10962247.2018.1459956.

Goh, C.C., Kamarudin, L.M., Zakaria, A., Nishizaki, H., Ramli, N., Mao, X., Zakaria, S.M.M.S., Kanagaraj, E., Sukor, A.S.A., Elham, M.F., 2021. Real-time in-vehicle air quality monitoring system using machine learning prediction algorithm. Sensors 21 (15), 4956. http://dx.doi.org/10.3390/s21154956.

Guo, Z., Jing, X., Ling, Y., Yang, Y., Jing, N., Yuan, R., Liu, Y., 2024. Optimized air quality management based on air quality index prediction and air pollutants identification in representative cities in China. Sci. Rep. 14 (1), http://dx.doi.org/10.1038/s41598-024-68972-w.

Gupta, N.S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., Arulkumaran, G., 2023. Prediction of air quality index using machine learning techniques: A comparative analysis. J. Environ. Public Heal. 2023, 1–26. http://dx.doi.org/10.1155/2023/4916267.

Hasan, K., Rahman, M., Akhter, M., Mohinuzzaman, M., Kayes, I., Rahman, S., 2024. A new dynamic approach using data-driven and machine learning models for mega city Dhaka's particulate matter forecasting. Environ. Pollut. Manag. http://dx.doi.org/10.1016/j.epm.2024.11.005.

Hulkkonen, M., Lipponen, A., Mielonen, T., Kokkola, H., Prisle, N.L., 2022. Changes in urban air pollution after a shift in anthropogenic activity analysed with ensemble learning, competitive learning and unsupervised clustering. Atmos. Pollut. Res. 13 (5), 101393. http://dx.doi.org/10.1016/j.apr.2022.101393.

Ionascu, M.-E., Marcu, M., Bogdan, R., Darie, M., 2024. Calibration of NO, SO2, and PM using airify: A low-cost sensor cluster for air quality monitoring. Atmos. Environ. 120841. http://dx.doi.org/10.1016/j.atmosenv.2024.120841.

Iskandaryan, D., Ramos, F., Trilles, S., 2023. Graph neural network for air quality prediction: A case study in madrid. IEEE Access 11, 2729–2742. http://dx.doi.org/10.1109/access.2023.3234214.

Jiménez-Navarro, M.J., Lovrić, M., Kecorius, S., Nyarko, E.K., Martínez-Ballesteros, M., 2024. Explainable deep learning on multi-target time series forecasting: an air pollution use case. Results Eng. 24, 103290. http://dx.doi.org/10.1016/j.rineng.2024.103290.

Kalantari, E., Gholami, H., Malakooti, H., Nafarzadegan, A.R., Moosavi, V., 2024. Machine learning for air quality index (AQI) forecasting: shallow learning or deep learning? Environ. Sci. Pollut. Res. 31 (54), 62962–62982. http://dx.doi.org/10.1007/s11356-024-35404-1.

Karimian, H., Li, Q., Wu, C., Qi, Y., Mo, Y., Chen, G., Zhang, X., Sachdeva, S., 2019. Evaluation of different machine learning approaches to forecasting PM2.5 mass concentrations. Aerosol Air Qual. Res. 19 (6), 1400–1410. http://dx.doi.org/10.4209/aaqr.2018.12.0450.

Kataria, A., Puri, V., 2022. AI- and IoT-based hybrid model for air quality prediction in a smart city with network assistance. IET Networks 11 (6), 221–233. http://dx.doi.org/10.1049/ntw2.12053.

Kim, H.S.H., 2023. Deep insight into urban air quality utilizing neural networks for enhanced prediction in Korean cities where factories and ecosystem environments coexists. Int. J. Recent. Innov. Trends Computing Commun. 11 (9), 1003–1009. http://dx.doi.org/10.17762/ijritcc.v11i9.8991.

Kim, S.H., Moon, H.J., 2023. The performance of reinforcement learning for indoor climate control devices according to the level of outdoor air particulate matters. Buildings 13 (12), 3062. http://dx.doi.org/10.3390/buildings13123062.

Kumar, K., Pande, B.P., 2022. Air pollution prediction with machine learning: a case study of Indian cities. Int. J. Environ. Sci. Technol. 20 (5), 5333–5348. http://dx.doi.org/10.1007/s13762-022-04241-5.

Kunak, 2023. Guide about urban pollution. https://kunakair.com/guide-about-urban-pollution/.

Laukkarinen, A., Vinha, J., 2024. Long-term prediction of hourly indoor air temperature using machine learning. Energy Build. 114972. http://dx.doi.org/10.1016/j.enbuild.2024.114972.

Li, L., Wang, J., Franklin, M., Yin, Q., Wu, J., Camps-Valls, G., Zhu, Z., Wang, C., Ge, Y., Reichstein, M., 2023. Improving air quality assessment using physics-inspired deep graph learning. Npj Clim. Atmos. Sci. 6 (1), http://dx.doi.org/10.1038/s41612-023-00475-3.

Lin, C., Yousefi, S., Kahoro, E., Karisani, P., Liang, D., Sarnat, J., Agichtein, E., 2022. Detecting elevated air pollution levels by monitoring web search queries: Algorithm development and validation. JMIR Form. Res. 6 (12), e23422. http://dx.doi.org/10.2196/23422.

Liu, B., Bryson, J.R., Sevinc, D., Cole, M.A., Elliott, R.J.R., Bartington, S.E., Bloss, W.J., Shi, Z., 2023. Assessing the impacts of birmingham's clean air zone on air quality: Estimates from a machine learning and synthetic control approach. Environ. Resour. Econ. 86 (1–2), 203–231. http://dx.doi.org/10.1007/s10640-023-00794-2.

Liu, Q., Cui, B., Liu, Z., 2024. Air quality class prediction using machine learning methods based on monitoring data and secondary modeling. Atmosphere 15 (5), 553. http://dx.doi.org/10.3390/atmos15050553.

Ly, H.-B., Le, L.M., Van Phi, L., Phan, V.-H., Tran, V.Q., Pham, B.T., Le, T.-T., Derrible, S., 2019. Development of an AI model to measure traffic air pollution from multisensor and weather data. Sensors 19 (22), 4941. http://dx.doi.org/10.3390/s19224941.

Lyu, Y., Ju, Q., Lv, F., Feng, J., Pang, X., Li, X., 2022. Spatiotemporal variations of air pollutants and ozone prediction using machine learning algorithms in the Beijing-Tianjin-Hebei region from 2014 to 2021. Environ. Pollut. 306, 119420. http://dx.doi.org/10.1016/j.envpol.2022.119420.

Ma, X., Chen, T., Ge, R., Xv, F., Cui, C., Li, J., 2023. Prediction of PM2.5 concentration using spatiotemporal data with machine learning models. Atmosphere 14 (10), 1517. http://dx.doi.org/10.3390/atmos14101517.

Mahmoud, A., Mohammed, A., 2024. Leveraging hybrid deep learning models for enhanced multivariate time series forecasting. Neural Process. Lett. 56 (5), http://dx.doi.org/10.1007/s11063-024-11656-3.

Mampitiya, L., Rathnayake, N., Hoshino, Y., Rathnayake, U., 2023. Forecasting PM10 levels in Sri Lanka: A comparative analysis of machine learning models PM10. J. Hazard. Mater. Adv. 13, 100395. http://dx.doi.org/10.1016/j.hazadv.2023.100395.

Miskell, G., Pattinson, W., Weissert, L., Williams, D., 2019. Forecasting short-term peak concentrations from a network of air quality instruments measuring PM2.5 using boosted gradient machine models. J. Environ. Manag. 242, 56–64. http://dx.doi.org/10.1016/j.jenvman.2019.04.010.

Morapedi, T.D., Obagbuwa, I.C., 2023. Air pollution particulate matter (PM2.5) prediction in South African cities using machine learning techniques. Front. Artif. Intell. 6, http://dx.doi.org/10.3389/frai.2023.1230087.

Moursi, A.S.A.E.A., Shouman, M., Hemdan, E.E.-D., El-Fishawy, N., 2019. PM2.5 concentration prediction for air pollution using machine learning algorithms. Menoufia J. Electron. Eng. Res. 28 (1), 349–354. http://dx.doi.org/10.21608/mjeer.2019.67375.

Mulomba, C., Choi, H., 2024a. Air quality forecasting using machine learning: A global perspective with relevance to low-resource settings. SSRN Electron. J. http://dx.doi.org/10.2139/ssrn.4686946.

Munir, S., Luo, Z., Dixon, T., Manla, G., Francis, D., Chen, H., Liu, Y., 2022. The impact of smart traffic interventions on roadside air quality employing machine learning approaches. Transp. Res. Part D Transp. Environ. 110, 103408. http://dx.doi.org/10.1016/j.trd.2022.103408.

Natarajan, S.K., Shanmurthy, P., Arockiam, D., Balusamy, B., Selvarajan, S., 2024. Optimized machine learning model for air quality index prediction in major cities in India. Sci. Rep. 14 (1), http://dx.doi.org/10.1038/s41598-024-54807-1.

Nguyen, P.H., Dao, N.K., Nguyen, L.S.P., 2024a. Development of machine learning and deep learning prediction models for PM2.5 in Ho Chi Minh City, Vietnam. Atmosphere 15 (10), 1163. http://dx.doi.org/10.3390/atmos15101163.

Nguyen, H.A.D., Le, T.H., Ha, Q.P., Duc, H., Azzi, M., 2024b. Particulate matter monitoring and forecast with integrated low-cost sensor networks and air-quality monitoring stations. E3S Web Conf. 496, 04001. http://dx.doi.org/10.1051/e3sconf/202449604001.

Nguyen, A.T., Pham, D.H., Oo, B.L., Ahn, Y., Lim, B.T.H., 2024d. Predicting air quality index using attention hybrid deep learning and quantum-inspired particle swarm optimization. J. Big Data 11 (1), http://dx.doi.org/10.1186/s40537-024-00926-5.

Oransi, 2023. Air quality in rural areas. https://oransi.com/blogs/how-it-works/air-quality-rural-areas/.

Oseni, K., Balogun, H., Sidhu, K.K., 2024. Predictive modelling of air quality index (AQI) across diverse cities and states of India using machine learning: Investigating the influence of punjab's stubble burning on AQI variability. SSRN Electron. J. http://dx.doi.org/10.2139/ssrn.4790794.

Pandey, A., Brauer, M., Cropper, M.L., Balakrishnan, K., Mathur, et al., 2020. Health and economic impact of air pollution in the states of India: the global burden of disease study 2019. Lancet Planet. Heal. 5 (1), e25–e38. http://dx.doi.org/10.1016/s2542-5196(20)30298-9.

Park, J., Seo, Y., Cho, J., 2023. Unsupervised outlier detection for time-series data of indoor air quality using LSTM autoencoder with ensemble method. J. Big Data 10 (1), http://dx.doi.org/10.1186/s40537-023-00746-z.

Pazhanivel, D.B., Velu, A.N., Palaniappan, B.S., 2024. Design and enhancement of a fog-enabled air quality monitoring and prediction system: An optimized lightweight deep learning model for a smart fog environmental gateway. Sensors 24 (15), 5069. http://dx.doi.org/10.3390/s24155069.

Qiu, M., Zigler, C., Selin, N.E., 2022. Statistical and machine learning methods for evaluating trends in air quality under changing meteorological conditions. Atmos. Chem. Phys. 22 (16), 10551–10566. http://dx.doi.org/10.5194/acp-22-10551-2022.

Rakholia, R., Le, Q., Ho, B.Q., Vu, K., Carbajo, R.S., 2023. Multi-output machine learning model for regional air pollution forecasting in Ho Chi Minh City, Vietnam. Environ. Int. 173, 107848. http://dx.doi.org/10.1016/j.envint.2023.107848.

Rakholia, R., Le, Q., Vu, K., Ho, B.Q., Carbajo, R.S., 2024. Accurate PM2.5 urban air pollution forecasting using multivariate ensemble learning accounting for evolving target distributions. Chemosphere 364, 143097. http://dx.doi.org/10.1016/j.chemosphere.2024.143097.

Ramirez-Alcocer, U.M., Tello-Leal, E., Macías-Hernández, B.A., Hernandez-Resendiz, J.D., 2022. Data-driven prediction of COVID-19 daily new cases through a hybrid approach of machine learning unsupervised and deep learning. Atmosphere 13 (8), 1205. http://dx.doi.org/10.3390/atmos13081205.

Rautela, K.S., Goyal, M.K., 2024. Transforming air pollution management in India with AI and machine learning technologies. Sci. Rep. 14 (1), http://dx.doi.org/10.1038/s41598-024-71269-7.

Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., Sonne, C., 2023. Air quality prediction by machine learning models: A predictive study on the indian coastal city of visakhapatnam. Chemosphere 338, 139518. http://dx.doi.org/10.1016/j.chemosphere.2023.139518.

Rescio, G., Manni, A., Caroppo, A., Carluccio, A.M., Siciliano, P., Leone, A., 2023. Multi-sensor platform for predictive air quality monitoring. Sensors 23 (11), 5139. http://dx.doi.org/10.3390/s23115139.

Ritchie, H., Roser, M., 2024. Air pollution. https://ourworldindata.org/air-pollution.

Rollo, F., Bachechi, C., Po, L., 2023. Anomaly detection and repairing for improving air quality monitoring. Sensors 23 (2), 640. http://dx.doi.org/10.3390/s23020640.

Rovira, J., Paredes-Ahumada, J., Barceló-Ordinas, J., García-Vidal, J., Reche, C., Sola, Y., Fung, P., Petäjä, T., Hussein, T., Viana, M., 2022. Non-linear models for black carbon exposure modelling using air pollution datasets. Environ. Res. 212, 113269. http://dx.doi.org/10.1016/j.envres.2022.113269.

Shankar, L., Arasu, K., 2023. Deep learning techniques for air quality prediction: A focus on PM2.5 and periodicity. Migration Lett. 20 (S13), 468–484. http://dx.doi.org/10.59670/ml.v20is13.6477.

Sharma, E., Deo, R.C., Prasad, R., Parisi, A.V., Raj, N., 2020. Deep air quality forecasts: Suspended particulate matter modeling with convolutional neural and long short-term memory networks. IEEE Access 8, 209503–209516. http://dx.doi.org/10.1109/access.2020.3039002.

Shetty, C., Seema, S., Sowmya, B.J., Nandalike, R., Supreeth, S., P, D., S, R., Y, V., Ranjan, R., Goud, V., 2024. A machine learning approach for environmental assessment on air quality and mitigation strategy. J. Eng. 2024, 1–16. http://dx.doi.org/10.1155/2024/2893021.

Singh, R.V., Singh, A.P., Kumar, S., 2024. Hybrid machine learning models for fine-grained air quality forecasting. Int. J. Multidiscip. Res. 6 (2), http://dx.doi.org/10.36948/ijfmr.2024.v06i02.18383.

Sridhar, K., Radhakrishnan, P., Swapna, G., Kesavamoorthy, R., Pallavi, L., Thiagarajan, R., 2022. A modular IOT sensing platform using hybrid learning ability for air quality prediction. Meas. Sensors 25, 100609. http://dx.doi.org/10.1016/j.measen.2022.100609.

Tamas, W., Notton, G., Paoli, C., Nivet, M.-L., Voyant, C., 2015. Hybridization of air quality forecasting models using machine learning and clustering: An original approach to detect pollutant peaks. Aerosol Air Qual. Res. 16 (2), 405–416. http://dx.doi.org/10.4209/aaqr.2015.03.0193.

Ugwu, C.N., Eze, V.H.U., Ogenyi, F.C., 2024. Urban air quality and machine learning. Res. Output J. Eng. Sci. Res. 3 (1), 101–104.

Vajs, I., Drajic, D., Cica, Z., 2023. Data-driven machine learning calibration propagation in a hybrid sensor network for air quality monitoring. Sensors 23 (5), 2815. http://dx.doi.org/10.3390/s23052815.

Vasudevan, P., Ekambaram, C., 2024. HYAQP: A hybrid meta-heuristic optimization model for air quality prediction using unsupervised machine learning paradigms. Int. Arab. J. Inf. Technol. 21 (5), http://dx.doi.org/10.34028/iajit/21/5/15.

Vieru, M.-C., Cărbureanu, M., 2024. Machine learning methods applied in air quality prediction. Rom. J. Pet. Gas Technol. 5 (76), 5–18. http://dx.doi.org/10.51865/jpgt.2024.01.01.

Vm, M., Gh, S.G., Kamalapurkar, S., 2020. Air pollution prediction using machine learning supervised learning approach. Int. J. Sci. Technol. Res. 9 (4), 118–123, URL https://www.ijstr.org/final-print/apr2020/Air-Pollution-Prediction-Using-Machine-Learning-Supervised-Learning-Approach.pdf.

Wang, A., Machida, Y., deSouza, P., Mora, S., Duhl, T., Hudda, N., Durant, J.L., Duarte, F., Ratti, C., 2023. Leveraging machine learning algorithms to advance low-cost air sensor calibration in stationary and mobile settings. Atmos. Environ. 301, 119692. http://dx.doi.org/10.1016/j.atmosenv.2023.119692.

Wang, X., Zhang, S., Chen, Y., He, L., Ren, Y., Zhang, Z., Li, J., Zhang, S., 2024. Air quality forecasting using a spatiotemporal hybrid deep learning model based on VMD–GAT–BiLSTM. Sci. Rep. 14 (1), http://dx.doi.org/10.1038/s41598-024-68874-x.

Waqas, N.H.M.S., Naz, N.T., Shahid, N.K., Ahmad, N.M.B., Yaqoob, N.A., Ahmad, N.M.B., 2024. Time series forecasting of air quality index in lahore : A machine learning perspective with facebook prophet model. Int. J. Sci. Res. Sci. Technol. 11 (5), 219–226. http://dx.doi.org/10.32628/ijsrst2411598.

Wei, P., Hao, S., Shi, Y., Anand, A., Wang, Y., Chu, M., Ning, Z., 2024. Combining google traffic map with deep learning model to predict street-level traffic-related air pollutants in a complex urban environment. Environ. Int. 191, 108992. http://dx.doi.org/10.1016/j.envint.2024.108992.

World Health Organization, 2021a. Air Quality Guidelines - Update 2021. WHO Regional Office for Europe, Copenhagen, Denmark.

World Health Organization, 2021b. New 2021 WHO air quality guideline limits. https://www.breeze-technologies.de/blog/new-2021-who-air-quality-guideline-limits/.

World Health Organization, 2021c. New WHO global air quality guidelines recommend new air quality levels to protect human health. WHO News.

World Health Organization, 2024. Air pollution - global health observatory data. https://www.who.int/data/gho/data/themes/air-pollution.

Wu, Y.-C., Shiledar, A., Li, Y.-C., Wong, J., Feng, S., Chen, X., Chen, C., Jin, K., Janamian, S., Yang, Z., Ballard, Z.S., Göröcs, Z., Feizi, A., Ozcan, A., 2017. Air quality monitoring using mobile microscopy and machine learning. Light. Sci. Appl. 6 (9), e17046. http://dx.doi.org/10.1038/lsa.2017.46.

Xayasouk, T., Lee, H., Lee, G., 2020. Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models. Sustainability 12 (6), 2570. http://dx.doi.org/10.3390/su12062570.

Yang, W., Li, H., Wang, J., Ma, H., 2024. Spatio-temporal feature interpretable model for air quality forecasting. Ecol. Indic. 167, 112609. http://dx.doi.org/10.1016/j.ecolind.2024.112609.

Yasmin, F., Hassan, M.M., Hasan, M., Zaman, S., Angon, J.H., Bairagi, A.K., Changchun, Y., 2023. AQIPred: A hybrid model for high precision time specific forecasting of air quality index with cluster analysis. Human-Centric Intell. Syst. 3 (3), 275–295. http://dx.doi.org/10.1007/s44230-023-00039-x.

Zhao, Z., Wu, J., Cai, F., Zhang, S., Wang, Y.-G., 2023. A hybrid deep learning framework for air quality prediction with spatial autocorrelation during the COVID-19 pandemic. Sci. Rep. 13 (1), http://dx.doi.org/10.1038/s41598-023-28287-8.

Zhivkov, P., 2021. Optimization and evaluation of calibration for low-cost air quality sensors: Supervised and unsupervised machine learning models. Ann. Comput. Sci. Inf. Syst. 25, 255–258. http://dx.doi.org/10.15439/2021f95.

Zhu, Q., Lee, D., Stoner, O., 2024. A comparison of statistical and machine learning models for spatio-temporal prediction of ambient air pollutant concentrations in Scotland. Environ. Ecol. Stat. http://dx.doi.org/10.1007/s10651-024-00635-5.