Artificial Intelligence Review (2025) 58:275
https://doi.org/10.1007/s10462-025-11277-9

Advancements in air quality monitoring: a systematic review of IoT-based air quality monitoring and AI technologies

Antony Garcia[1,4] · **Yessica Saez**[1,3] · **Itamar Harris**[2] · **Xinming Huang**[4] · **Edwin Collado**[1,3]

Accepted: 20 May 2025 / Published online: 11 June 2025

Abstract
Air quality monitoring is a critical component of environmental management, especially in developing countries where pollution levels are a growing concern. This review systematically analyzes recent advances in Internet-of-Things (IoT)-based air quality monitoring systems and highlights the impact of artificial intelligence (AI) technologies. A comprehensive selection process was used to identify 220 relevant studies using databases such as Google Scholar, Scopus, and ResearchRabbit. Following a selection process based on specific eligibility criteria, a total of 147 studies were chosen for a detailed analysis. These studies present notable advances in the application of artificial intelligence to improve data accuracy, predictive capabilities, and real-time analysis in air quality monitoring systems. A key contribution of this review is the proposal of a classification framework for AI techniques in air quality monitoring, organized into five main application areas: data imputation, sensor calibration, anomaly detection, air quality index (AQI) estimation, and short-term forecasting. The review takes an in-depth look at the different uses of these technologies in both urban and industrial settings, presenting successful case studies that showcase their effectiveness in addressing pressing air quality issues. Additionally, the paper identifies research gaps in the literature, particularly related to data quality, system scalability, and integration challenges in AI-driven IoT systems. The insights provided aim to guide researchers and practitioners in selecting appropriate AI techniques and system architectures, inform the design of more reliable and scalable air quality monitoring frameworks, and support future efforts to mitigate air pollution through data-driven decision-making.

Keywords Air quality monitoring · Artificial intelligence · Deep learning · Internet of things · Machine learning · Predictive modeling · Sensor calibration

Extended author information available on the last page of the article

1 Introduction

Air quality has become a critical concern in recent years due to its impact on public health and the environment. Poor air quality is associated with increased morbidity and mortality, with a large proportion of urban populations exposed to air that does not meet health standards (Kelly and Fussell 2015). Climate change exacerbates this issue, affecting the levels of key pollutants such as ozone, carbon dioxide, nitrogen dioxide, sulfur dioxide, and particle matter, all of which have serious health implications (Hassan et al. 2016). Furthermore, studies have shown that air pollution is responsible for respiratory problems, including bronchitis, pneumonia, emphysema, asthma, and lung cancer (Sunyer et al. 2006; Kravchenko et al. 2014; Sarnat and Holguin 2007; Buonanno et al. 2017).

To address this problem, several advanced technologies have been applied to monitor and improve air quality. Machine learning (ML) and deep learning (DL) algorithms analyze extensive environmental data, leading to more accurate predictions and real-time monitoring of air pollution levels (Sai et al. 2019; William et al. 2023). AI-driven Internet of Things (IoT) systems facilitate the continuous monitoring of air quality, providing timely data to inform public health interventions and policy decisions (Bharathi et al. 2022; Tran et al. 2022; Marzouk and Atef 2022). These technologies have been implemented in various applications, from monitoring indoor air quality to predicting pollution patterns in urban areas, thus helping mitigate the adverse effects of air pollution on human health and the environment.

In the field of air quality monitoring, DL techniques have been used to address various challenges such as missing data, noise, and the need for accurate predictions. Techniques such as autoencoders and generative adversarial networks (GANs) are used to impute missing values in datasets, ensuring the integrity of the data used for analysis (Kim et al. 2021; Wu et al. 2022). Convolutional neural networks (CNN) and recurrent neural networks (RNN), particularly long-short-term memory networks (LSTM), have been applied to predict air quality levels by capturing spatial and temporal dependencies in environmental data (Mitreska Jovanovska et al. 2023; Osman et al. 2024). DL models are also robust against sensor data noise, enabling reliable monitoring and improving sensor calibration and noise mitigation (Zimmerman et al. 2018; Du et al. 2021). These models can be fine-tuned to adapt to different types of pollutants and environmental conditions, making them versatile tools in the continuous battle against air pollution.

Compared to traditional air quality monitoring methods, which often rely on expensive, stationary equipment and require manual data retrieval and processing, AI-based IoT solutions offer several advantages. These include the ability to provide real-time, continuous data collection; lower deployment and maintenance costs through the use of low-cost sensors; improved scalability for wide-area monitoring; and enhanced predictive and analytical capabilities through machine learning. These benefits make AI-IoT systems particularly attractive for addressing the growing demands of modern air quality management, especially in resource-constrained environments.

Although there have been advances in air quality monitoring and management, several challenges remain. Most global regions have sparse monitoring networks and inconsistent data quality, which hinders accurate air quality assessment (Xu and Zhu 2016). Studies have mentioned the importance of multiparameter approaches and IoT-based systems to improve urban monitoring precision and coverage (Marinov et al. 2016), but technical challenges

such as sensor calibration, limited operational life, and integration into large networks further complicate the efficacy of continuous and real-time monitoring (Maag et al. 2018; Bílek et al. 2021).

Recent reviews have provided valuable syntheses of AI-based approaches to air quality forecasting, particularly emphasizing advancements in deep learning for spatiotemporal modeling (Garbagna et al. 2025; Kaginalkar et al. 2021). However, these studies primarily focus on predictive techniques and often lack a unified framework that categorizes the diverse functions AI can serve in this domain. In contrast, the present review introduces a structured classification system that organizes AI applications into five core functional areas: data imputation, sensor calibration, anomaly detection, AQI estimation, and short-term forecasting. This taxonomy, illustrated in Fig. 1, guides the logical flow and thematic organization of the review, providing a cohesive structure for analyzing the literature.

The review begins with a comprehensive overview of recent developments in air quality monitoring, with particular attention to IoT-based systems and the integration of AI technologies. It then outlines the methodological foundation based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyzes (PRISMA) guidelines, detailing the criteria for literature selection, screening processes, and formulation of research questions. The proposed taxonomy is then applied to explore how AI and IoT platforms converge to enable real-time, scalable, and accurate air quality monitoring solutions.

Each functional domain within the taxonomy is analyzed in depth, supported by case studies and examples that demonstrate successful applications in both urban and industrial environments. Additionally, the review addresses critical challenges associated with implementing AI and IoT-based systems, including issues related to data quality, sensor accuracy, computational limitations, and scalability. By doing so, it not only highlights the current

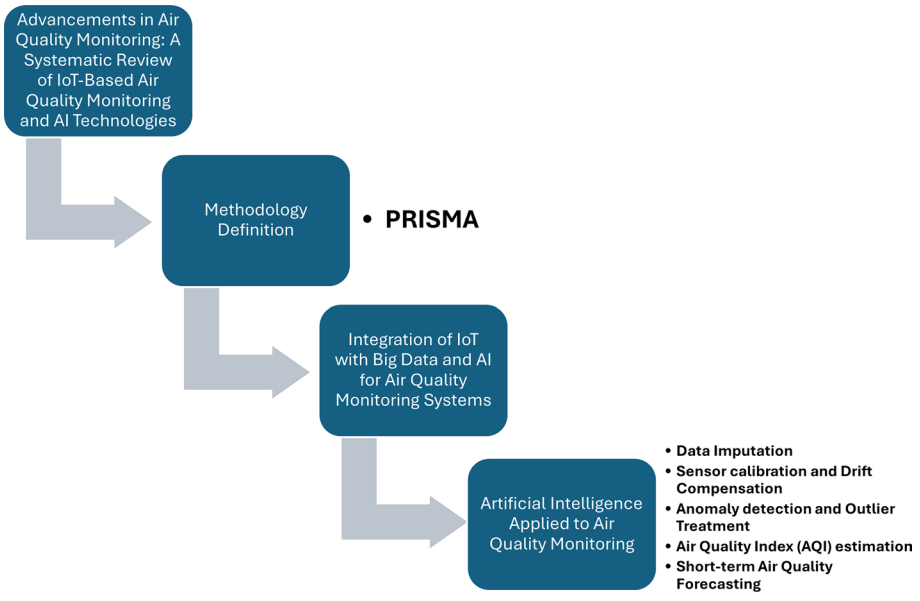


Fig. 1 Taxonomy of this study - advancements in air quality monitoring: a systematic review of IoT-based air quality monitoring and AI technologies

state of the field but also identifies emerging technologies and gaps in the literature that present opportunities for future research.

Ultimately, this review offers a novel contribution by systematically categorizing AI techniques according to their functional roles in air quality monitoring and by integrating this taxonomy with a broader discussion on technological implementation and practical challenges. This dual focus allows for a comprehensive and actionable synthesis that advances understanding and supports the development of more robust and intelligent air quality management systems.

2 Methodology

This survey aims to examine the application of AI technologies in IoT-based air quality monitoring systems. The literature search was conducted using the tools Publish or Perish 8 and ResearchRabbit. Publish or Perish was used to source literature from reputable databases such as Google Scholar and SCOPUS. ResearchRabbit, which provides a visual representation of connections between articles, was used to identify potentially missing relevant articles that may not have been included in the initial keyword-based search, allowing for a more interconnected review of the literature.

Specific search keywords were used to identify the most recent advances in the field, including terms such as "weather stations", "Internet of Things", "air quality monitoring station", "air quality index estimation", "air quality monitoring", as well as related variations such as “environmental monitoring,” “pollution tracking,” “atmospheric monitoring”, "machine learning", "deep learning", and "artificial intelligence". These terms were combined to ensure a comprehensive and inclusive search strategy.

Through these tools, more than 266 research articles published between January 2016 and December 2024 were initially identified. After applying the selection criteria, 147 articles were included in the final review. Although the focus was primarily on recent developments, papers from the 1990s and 2000s were included to provide historical context on early developments of weather monitoring stations.

The study follows the PRSIMA guidelines (see Fig. 2) to ensure methodological rigor and transparency in the selection and evaluation of relevant literature.

2.1 Research questions

Following the PRISMA guidelines, this review was driven by the following research questions:

- **Research question 1**: What architectures have been used to build IoT-based air quality monitoring stations and what key features have been considered in their design?
- **Research question 2**: What data-related problems have been identified in IoT-based air quality monitoring?
- **Research question 3**: How have these data-related problems been addressed using AI techniques, such as ML and DL?
- **Research question 4**: What specific ML and DL models have been used to address data-related problems in IoT-based air quality monitoring?

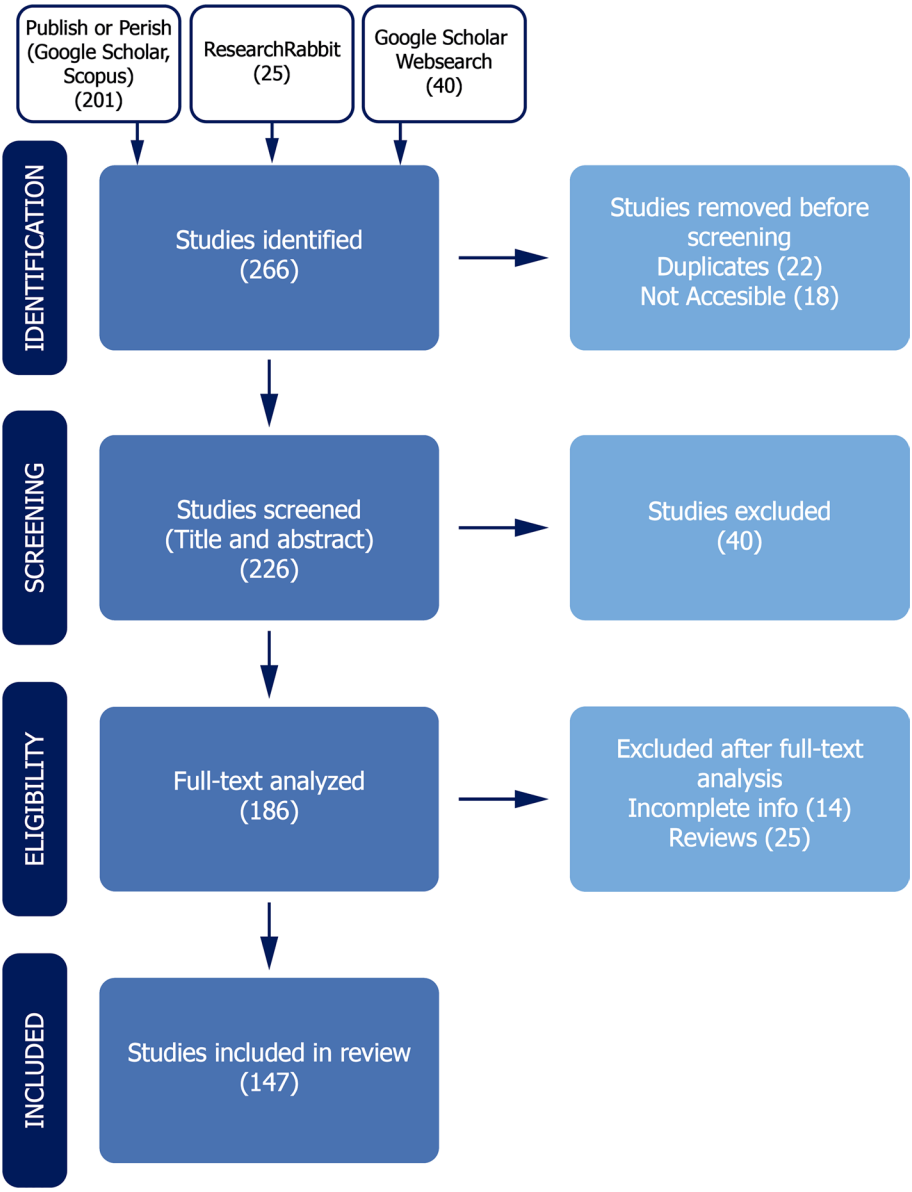


Fig. 2 Flow map of the systematic review process based on PRISMA (Page et al. 2021)

- **Research question 5**: What challenges remain in the implementation and scalability of AI-driven IoT-based air quality monitoring systems?

To ensure that the research questions were addressed with high-quality and relevant evidence, a rigorous screening process was applied, as outlined in the PRISMA flow diagram (Fig. 2). Each decision point in the selection process-ranging from title and abstract screen-

ing to full-text assessment-was guided by the eligibility criteria defined in Section 2.2. Studies were excluded primarily due to insufficient technical detail, lack of AI integration, or absence of peer review. The final pool of included articles reflects those that provided substantial insights into IoT architectures, data-related challenges, and the application of AI techniques in air quality monitoring, thereby aligning directly with the scope and objectives of this review.

2.2 Data collection and selection criteria

To ensure relevance and quality, the articles selected for review met the following inclusion criteria:

- Published between January 2016 and December 2024 in peer-reviewed journals or conferences.
- Articles focused on IoT-based air quality monitoring systems that incorporate various sensor networks.
- Studies that provide detailed information on hardware and software implementations relevant to IoT-based air quality monitoring.
- Studies that discuss AI applications in IoT-based air quality monitoring.
- Studies that used ML or DL techniques to address challenges related to data quality, processing, and interpretation.

Exclusion criteria included:

- Preprints or articles lacking detailed performance metrics.
- Studies focusing on non-IoT-based monitoring systems.
- Redundant or duplicate records across databases.
- Review papers, as they do not contribute original experimental findings.
- Articles that have never been cited by other research papers, suggesting a limited academic impact.
- Studies lacking in-depth discussions on the integration of AI with IoT-based air quality monitoring systems.

To minimize selection bias, a dual reviewer strategy was employed throughout the article screening process. Two authors independently assessed titles, abstracts, and full texts based on predefined inclusion and exclusion criteria. In cases of disagreement, discussions were held until a consensus was reached. This process helped to ensure objectivity and consistency in the selection of studies. Additionally, the use of ResearchRabbit enabled the identification of influential papers that may not have been retrieved through standard keyword-based searches, thus reducing the risk of omitting relevant literature due to search term limitations. Finally, the exclusion of non-peer-reviewed articles, uncited articles, and duplicate entries further contributed to the rigor and reliability of the reviewed literature. To ensure consistency, a pilot screening of 20 articles was conducted to refine the eligibility criteria, and inter-reviewer agreement was discussed during initial and final selection stages.

Beyond inclusion/exclusion, each study was evaluated based on the depth of technical detail, the clarity of the methodology, and the relevance to the defined research questions.

Studies were also categorized according to their primary contribution to one or more of the five functional areas defined in our proposed classification framework: data imputation, sensor calibration, anomaly detection, AQI estimation, and short-term forecasting. This coding process allowed for a systematic synthesis of findings across studies and helped ensure consistency, transparency, and reproducibility in the analysis.

2.3 Bibliometric findings

The co-occurrence of keywords in the documents retrieved from the screened studies (see Fig. 2) was analyzed using VOSviewer software (version 1.6.18). Figure 3 illustrates the frequency of keyword occurrences, drawn from a pool of 1,724 keywords. Only those that appeared at least 10 times were included in the mapping, resulting in 41 distinct keywords. Links between keywords indicate frequent co-occurrence within the same documents, while closely related keywords are grouped into clusters. The software identified three clusters, led by the thematic topics *Internet of Things*, *Machine Learning*, and *Deep Learning*. In particular, terms such as *air quality*, *air pollution*, *particulate matter*, *air pollutant*, and *forecasting* were the most recurrent, represented by larger circles.

Beyond the frequency of keywords, the color gradient in Fig. 3 represents the average number of citations of articles associated with each keyword. Keywords related to advanced predictive modeling-such as *long short-term memory*, *forecasting*, and *deep learning*-as well as *air pollutants* and *air pollutant*, exhibited the highest average citation counts. This suggests that research that incorporates AI techniques in air quality analysis has attracted scientific attention and impact.

It is worth noting the prominence of both *machine learning* and *deep learning* as separate yet frequently co-occurring keywords. While *machine learning* represents a broad spectrum of data-driven modeling techniques, *deep learning* includes a more specialized subset

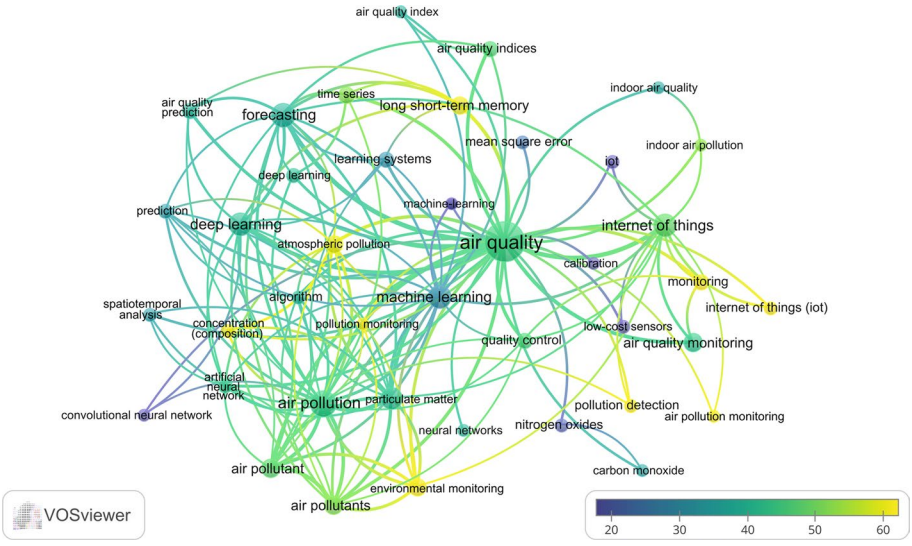


Fig. 3 Keyword co-occurrence network in literature on the application of AI technologies in IoT-based air quality monitoring systems. Colors indicate the average number of citations of publications using each keyword, with warmer colors (yellow) representing higher citation counts

focused on neural network architectures. The dominance of *deep learning*, particularly in conjunction with keywords such as *long short-term memory*, reflects a recent trend in the literature toward the use of advanced time series prediction models for air quality forecasting. This underscores the growing scientific interest in using complex AI methods for environmental monitoring tasks.

Following this analysis, a global mapping was conducted to assess the influence of research on the topic, recognizing contributions from 55 countries. Figure 4 highlights 21 countries with at least 3 publications related to the topic of interest. The findings suggest that China is the main contributor with a total of 61 documents, followed by India with 25, the United States with 13, and the United Kingdom with 12, indicating their numerous research efforts in this field.

Furthermore, Fig. 4 uses a color gradient to illustrate the average number of citations received by publications from each country. China leads with the highest total number of citations (2,045), followed by India (916), the United Kingdom (828), and the United States (631). Countries shaded with warmer colors indicate a higher citation average, showing the impact of each nation's research contributions. This visualization shows not only the volume of research but also the influence and recognition that these works have received within the academic community. Finally, the links in Fig. 4 represent the degree of collaboration between researchers from different countries, as reflected in the patterns of co-authorship in published articles.

The dominance of countries such as China, India, and the United States may be partly explained by their significant investments in smart city infrastructure, environmental data collection initiatives, and national strategies supporting AI integration in public health and environmental monitoring. Additionally, higher research funding, established academic networks, and the predominance of English-language publications likely contribute to increased visibility and citation rates. These factors collectively reflect regional research priorities and disparities in access to resources, which shape the global landscape of scholarly output in this domain. Finally, the links in Fig. 4 represent the degree of collaboration between researchers from different countries, as reflected in the patterns of co-authorship in published articles.

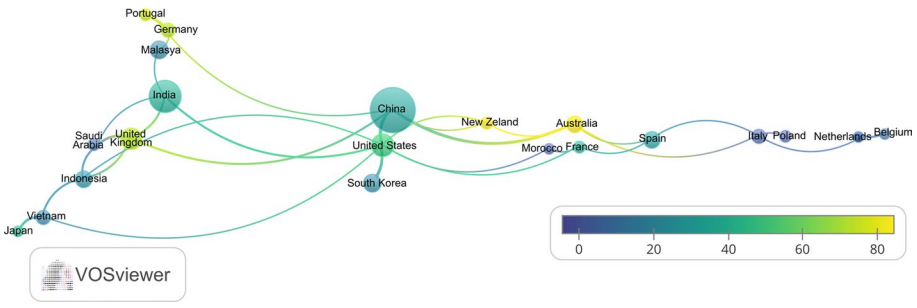


Fig. 4 Global co-authorship network map in the application AI technologies in IoT-based air quality monitoring systems research. Colors indicate the average number of citations of publications from each country, with higher citation counts represented by warmer colors

2.4 Article selection process

The article selection process was conducted in multiple stages to ensure the inclusion of the most relevant and high-quality studies:

- **Pilot screening:** An initial pilot screening of 20 articles was conducted to refine and align the eligibility criteria and to establish inter-reviewer agreement.
- **Initial screening:** Duplicate articles and non-peer-reviewed articles, as well as uncited studies, were excluded based on their titles and metadata to ensure rigor and relevance.
- **Title and abstract review:** Articles were assessed for their relevance to air quality monitoring and AI integration, based on title and abstract content.
- **Full-text analysis:** Eligible studies were carefully examined to extract relevant methodologies, results, and findings, ensuring that they met the inclusion criteria.

2.5 Analytical approach

The analytical approach involved extracting and analyzing key parameters from the selected studies to gain insights into various aspects of IoT-based air quality monitoring. The extracted parameters include the following:

- **Hardware and sensors:** Examination of the IoT sensors and data acquisition methods used in the studies, including sensor types, data transmission techniques, and sensor placement in indoor and outdoor environments.
- **Machine Learning and Deep Learning models:** Identification of the types of ML and DL models used, such as decision trees (DT), random forest (RF), support vector machines (SVM), CNN, LSTM, and transformer models, and their application to air quality monitoring.
- **Challenges and solutions:** Identification of common challenges such as data quality issues, missing values, sensor calibration, and the proposed solutions using ML and DL techniques.
- **Trends and future directions:** Analysis of emerging trends and future research directions in IoT-based air quality monitoring, including advances in AI, sensor technology, and data integration.

Data synthesis and comparison were performed to derive meaningful insights and identify patterns in the selected studies, which are further discussed in the subsequent sections.

2.6 Scope of the study

This review systematically examines the role of artificial intelligence in IoT-based air quality monitoring, covering literature published between January 2016 and December 2024. It introduces a structured classification framework that organizes AI applications into five core functional domains: data imputation, sensor calibration, anomaly detection, AQI estimation, and short-term forecasting. This taxonomy guides the thematic organization of the review and enables a cohesive analysis of how AI contributes across different layers of air quality monitoring systems.

In addition to exploring algorithmic advancements, the review considers practical aspects such as the deployment of indoor and outdoor monitoring stations, the use of microcontrollers, and real-time data communication. It addresses key challenges reported in the literature, including data noise, missing values, sensor drift, and the need for scalable and efficient models. The study also discusses the application of AI techniques such as CNN, LSTM, transformer, and hybrid models that adapt to dynamic environmental conditions.

By integrating this functional taxonomy with an analysis of implementation challenges and emerging technologies, the review provides a comprehensive and actionable synthesis that informs the development of intelligent, scalable, and reliable air quality monitoring systems.

2.7 Limitations

Although this review examines AI-driven air quality monitoring solutions in detail, certain limitations must be acknowledged. The selection of articles may introduce potential biases, as it relies on specific databases and search criteria that could exclude relevant studies. Emerging developments and methodologies, arising from the rapidly changing landscape of AI algorithms, could extend beyond the period examined, possibly impacting the applicability and importance of the results. Furthermore, the variability in data sources, sensor accuracy, and environmental conditions in different studies presents challenges in drawing generalized conclusions.

3 Integration of IoT with big data and AI for air quality monitoring systems

Historically, air quality monitoring was performed using stationary air quality stations, which have been deployed worldwide since the 1960s (Hůnová et al. 2004). Initially, these stations relied on analog devices and circuits before transitioning to electronic systems following advances in the field after the 1970s.

Early electronic systems were primarily based on microcontrollers or microprocessors (Mukaro et al. 1999; Sparks and Sumner 1984). Although effective for basic data collection, they had several limitations. One major disadvantage was their inability to provide real-time data, which restricted their usefulness for immediate public health responses and hampered timely policy decisions. In addition, data were often stored locally, requiring manual retrieval, leading to delays in analysis and reporting.

Communication between these early air quality stations and central monitoring facilities was based on wired protocols, such as serial RS232 communication (Mukaro et al. 1997). Although reliable, these wired connections limited the scalability and flexibility of monitoring networks. The need for physical connections restricted the placement of monitoring stations, often leading to gaps in coverage, particularly in large or remote areas.

In the early 2000s, the development of wireless communication protocols substantially improved air quality monitoring by enabling real-time data transmission to data centers. Advancements in microcontroller technology allowed real-time data transfer to computers (Kularatna et al. 2008), and these systems eventually gained the ability to communicate wirelessly with networks. Initially, this was achieved through radio frequency communica-

tion (Chung et al. 2006). Over time, various communication protocols were introduced, including Bluetooth (Yang and Li 2015), WiFi (Postolache et al. 2009; Martín-Garín et al. 2018; Dhingra et al. 2019; Pradityo and Surantha 2019; Sai et al. 2019; Marques and Pitarma 2019; Das et al. 2022; Purbakawaca et al. 2022), GPRS (Al-Ali et al. 2010; Zhao et al. 2019; Collado et al. 2024), GSM (Purbakawaca et al. 2022), and Zigbee (Ma et al. 2014), among others. The incorporation of wireless communication technologies improved the flexibility, scalability, and effectiveness of air quality monitoring, making the rapid and effective collection of environmental data possible.

With the increasing integration of cloud computing and big data analytics, air quality monitoring has completely shifted to the IoT scenario, enabling better data accessibility and real-time decision making. The introduction of new technologies such as LoRa (Zhao et al. 2019; Husein et al. 2019), LoRaWAN (Jabbar et al. 2022), MQTT (Kumar and Jasuja 2017), NB-IoT (Das et al. 2022; Zhao et al. 2019), and Bluetooth Low Energy (BLE) (Palomeque-Mangut et al. 2022) has improved these systems by offering low-power and long-range communication options.

The integration of IoT with big data and AI has transformed air quality monitoring from a simple data collection process to an intelligent decision-making system. IoT devices, such as low-cost sensors (LCS) and portable air quality monitors, continuously generate large amounts of data in real time. When integrated with AI algorithms, these data make it possible to use methods to improve pollution forecasting, source identification, and exposure evaluation, among other applications. Machine learning techniques have been used to calibrate sensor data, correct biases, and improve the accuracy of air quality forecasting models (Kaginalkar et al. 2021).

Air quality monitoring systems have increasingly adopted microcontrollers, particularly those of the Atmel family, which are programmable using the Arduino language. These microcontrollers are preferred for their ease of use, flexibility, and strong community support, making them a popular choice in the field (Kumar and Jasuja 2017; Benammar et al. 2018; Dhingra et al. 2019; Husein et al. 2019; Pradityo and Surantha 2019; Sai et al. 2019; Jabbar et al. 2022; Purbakawaca et al. 2022; Collado et al. 2024).

The ESP8266 microcontroller, which also supports programming in the Arduino language, is notable for its cost-effective wireless communication capabilities, including native support for WiFi and Bluetooth. This makes it a good choice for IoT-based air quality monitoring systems by improving their data collection and transmission capabilities (Martín-Garín et al. 2018; Marques and Pitarma 2019).

It is also common to find combinations of Arduino (Atmel) and ESP8266 in these systems, using the strengths of both platforms to create robust and scalable monitoring solutions (Purbakawaca et al. 2022; Das et al. 2022; Sai et al. 2019; Dhingra et al. 2019). Furthermore, the use of Raspberry Pi, often in conjunction with these microcontrollers, provides increased processing power, enabling more complex data processing and analysis directly at the edge (Kumar and Jasuja 2017; Pradityo and Surantha 2019).

Advances in microcontroller technology have enabled edge computing. This scenario allows for the execution of ML and even DL algorithms directly on embedded systems such as the microcontrollers mentioned above (Baller et al. 2021). By bringing computational tasks closer to the data source, edge computing reduces latency and decreases the reliance on centralized cloud processing. This has the potential to improve the autonomy of air qual-

ity monitoring nodes, allowing local data analysis, anomaly detection, and adaptive sensing before transmitting only critical information to the cloud (Idrees et al. 2018).

These systems can detect sensor failures, identify irregularities in environmental readings, and recognize sudden changes in data patterns without relying on continuous cloud connectivity (Abimannan et al. 2023). This capability is particularly valuable in air quality monitoring, where a wide range of environmental variables must be tracked, including particulate matter ($PM_{2.5}$ and PM_{10}), carbon monoxide (CO_2), nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2), volatile organic compounds (VOCs), humidity, temperature, and barometric pressure.

Although these systems often use low-cost sensors due to their affordability and ease of integration with microcontrollers, such sensors are susceptible to calibration drift, data inconsistencies, and environmental interference. Several studies have evaluated the reliability of low-cost sensors in air quality monitoring, providing evidence that these systems can deliver accurate data when appropriate calibration and data processing techniques are applied (Morawska et al. 2018; Castell et al. 2017; Rai et al. 2017; Badura et al. 2018). For applications where the appropriate hardware resources are available, edge computing has the potential to mitigate these challenges through real-time error detection, sensor calibration adjustments, and data filtering at the source, improving overall system reliability.

4 Artificial intelligence applied to air quality monitoring

Artificial intelligence has emerged as an important tool in air quality monitoring, providing new ways to analyze and predict pollution levels. With the ability to process large data sets and identify complex patterns, ML enhances various aspects of air quality management. It has been applied to predict pollutant concentrations (Wang et al. 2022), classify air quality levels (Idroes et al. 2023), identify pollution sources (Mishra et al. 2023), and optimize sensor networks (Ullo and Sinha 2020). Furthermore, ML supports data imputation, sensor calibration, and real-time alert systems, improving the reliability of air quality assessments (Rad et al. 2022). When integrated with IoT frameworks, ML enables autonomous monitoring systems, facilitating timely and accurate decision making in environmental management (Samie et al. 2019).

After analyzing the topics addressed in the surveyed articles, several research areas emerged as the main fields of application of ML and DL in air quality monitoring with IoT-based systems. These areas represent the most studied applications in the field:

1. Data imputation for handling missing values.
2. Sensor calibration and drift compensation.
3. Anomaly detection and outlier treatment.
4. Air quality index (AQI) estimation, typically using regression techniques.
5. Short-term air quality forecasting.

The following sections examine these areas in detail, focusing on methodologies, challenges, and recent developments that contribute to air quality management.

4.1 Introduction to machine learning and deep learning algorithms

This section provides a brief overview of the most commonly used ML and DL techniques applied to air quality monitoring, serving as a foundation for understanding the study results discussed later.

4.1.1 Machine learning algorithms

Machine learning, a branch of artificial intelligence, focuses on developing algorithms that can learn from data and make predictions. It includes a range of techniques such as supervised learning, unsupervised learning, and clustering. Common ML algorithms include DT, SVM, and RF. These methods are designed to detect patterns in data and make predictions based on those patterns.

Building on this foundation, we now present the most frequently adopted ML algorithms in the context of air quality monitoring.

4.1.2 K-nearest neighbors

This algorithm is a simple and effective supervised learning method used for classification and regression tasks. It works by identifying the *k*-nearest data points in the training set to a given input point and making predictions based on the majority class (for classification) or the average value (for regression) of those neighbors. K-nearest network (KNN) is useful in ML due to its ability to handle non-linear relationships and its simplicity in implementation (Peterson 2009).

However, it can be computationally expensive for large datasets, as it requires calculating distances between the input point and all training points. Additionally, KNN is sensitive to the choice of *k* and the distance metric used, which can affect its performance.

4.1.3 Support vector machines

Support vector machines are supervised learning algorithms used for classification and regression tasks. They work by finding the optimal hyperplane that separates data points from different classes in a high-dimensional space (Noble 2006). SVMs are particularly effective for high-dimensional datasets and can handle non-linear relationships through the use of kernel functions.

Kernel functions allow SVMs to transform the input data into a higher-dimensional space, making it easier to find a separating hyperplane. Common kernel functions include linear, polynomial, and radial basis function (RBF) kernels.

SVMs have limitations, including sensitivity to the selection of kernel functions and hyperparameters, which can significantly influence their performance. Furthermore, they can be computationally intensive for large datasets due to the need to solve a quadratic optimization problem.

4.1.4 Random forests

Random forests are an ensemble learning method that combines multiple decision trees to improve prediction accuracy (Breiman 2001). Each tree in the forest is trained on a random subset of the data, and the final prediction is made by aggregating the predictions of all trees using methods like majority voting for classification or averaging for regression.

RFs are robust to noise and can handle high-dimensional datasets with many features. They also provide feature importance scores, which can help identify the most relevant variables for prediction.

Its limitations include the potential for overfitting if the number of trees is too large and the difficulty in interpreting the model due to its ensemble nature. Additionally, RFs can be computationally expensive for large datasets, as they require training multiple decision trees.

4.1.5 Decision trees

Decision trees are supervised learning algorithms that are used for classification and regression tasks. They work by recursively partitioning the input space into smaller regions based on feature values, creating a tree-like structure where each internal node represents a feature, each branch represents a decision rule, and each leaf node represents a predicted outcome (Kotsiantis 2011).

DTs are easy to interpret and visualize, which makes them useful for understanding the relationships between features and outcomes. They can handle categorical and continuous data and are robust to noise. They can be used for classification and regression tasks.

Despite their simplicity and interpretability, DTs have limitations. They are highly sensitive to small changes in the data, which can lead to significant variations in the tree structure and predictions. Additionally, overly complex trees are prone to overfitting, capturing noise rather than the underlying patterns in the data. Pruning methods are often employed to mitigate these issues by removing unnecessary branches and enhancing the model's generalization capabilities.

4.1.6 DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is an unsupervised clustering algorithm that groups data points based on their density. It identifies clusters of varying shapes and sizes by examining the local density of data points and separating them from noise or outliers (Schubert et al. 2017).

DBSCAN is used in machine learning for tasks such as anomaly detection, spatial data analysis, and clustering of high-dimensional datasets. It is particularly effective for identifying clusters in noisy data and can handle clusters of different shapes and sizes.

Its limitations include the need to specify two parameters: the radius (epsilon) for neighborhood search and the minimum number of points required to form a cluster. The choice of these parameters can significantly affect the clustering results. Additionally, DBSCAN may struggle with clusters of varying densities, as it relies on a fixed density threshold.

4.1.7 Deep learning algorithms

Deep learning, a subfield of ML, uses neural networks with multiple layers to learn complex patterns from large datasets. Several DL architectures have been developed for this purpose, including CNN, autoencoders, LSTM networks, and transformers. These models are capable of learning hierarchical representations of data, enabling them to capture complex relationships and produce accurate predictions.

We now introduce the most commonly adopted DL algorithms in the context of air quality monitoring.

4.1.8 Artificial neural networks

Artificial neural networks (ANN) are a class of machine learning models inspired by the structure and function of the human brain. They consist of interconnected nodes (neurons) organized into layers, including an input layer, one or more hidden layers, and an output layer. ANNs are capable of learning complex patterns and relationships in data through a process called backpropagation, where the model adjusts its weights based on the error between predicted and actual outputs (Krogh 2008).

ANNs are widely used in various applications and are the building blocks of DL models. They can handle both structured and unstructured data, making them versatile for tasks such as classification, regression, and time series forecasting.

ANN has the ability to learn complex non-linear relationships in data, making them suitable for a wide range of applications. Their limitations include the need for large amounts of labeled data for training, the risk of overfitting if the model is too complex, and the difficulty in interpreting the learned representations. Additionally, ANNs can be computationally expensive to train, especially for deep architectures with many layers.

4.1.9 Convolutional neural networks

Convolutional neural networks are a specialized type of artificial neural network designed for processing structured grid data, such as images or time series. They consist of convolutional layers that apply filters to the input data, capturing local patterns and spatial hierarchies. CNNs are particularly effective for tasks such as image classification, object detection, and time series forecasting (LeCun et al. 1989).

CNN are the most widely used deep learning architectures for image and video analysis. They are designed to automatically learn spatial hierarchies of features from input data, making them particularly effective for tasks such as image classification, object detection, and segmentation (LeCun et al. 2015).

Its limitations include the need for large amounts of labeled data for training, the risk of overfitting if the model is too complex, and the difficulty in interpreting the learned representations. Additionally, CNNs can be computationally expensive to train, especially for deep architectures with many layers.

4.1.10 Autoencoders

Autoencoders are a type of neural network used for unsupervised learning tasks, such as dimensionality reduction, feature extraction, and anomaly detection. They consist of an encoder that compresses the input data into a lower-dimensional representation and a decoder that reconstructs the original data from this representation. Autoencoders can learn meaningful features from the data without requiring labeled examples.

Autoencoders are one of the most useful unsupervised learning techniques in DL, with several applications accross various domains. In the context of air quality monitoring, they can be used for tasks such as data denoising, anomaly detection, and feature extraction from complex datasets (Vincent et al. 2008). A workflow of an autoencoder used for data imputation in an air quality monitoring study can be seen in Fig. 5.

These architectures have some limitations, including the requirement for substantial amounts of data for effective training and the potential risk of overfitting when the model is overly complex. They are also sensitive to the selection of architecture and hyperparameters, which can significantly influence their performance. Furthermore, interpreting the learned representations can be challenging, especially in high-dimensional datasets.

4.1.11 Long short-term memory networks

Long short-term memory networks are a type of recurrent neural network designed to handle sequential data and capture long-range dependencie (Hochreiter and Schmidhuber 1997) s. They consist of memory cells that can store information over time, allowing them to learn patterns in time series data. LSTMs are widely used for tasks such as natural language processing, speech recognition, and time series forecasting.

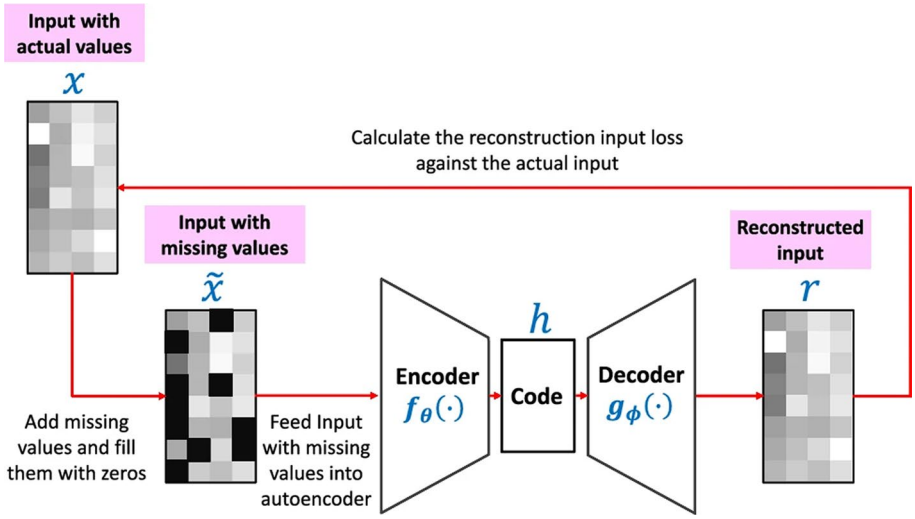


Fig. 5 Autoencoder-based approach for imputing missing data. The input with missing values is processed through an encoder-decoder structure, where the encoder compresses the input into a latent representation, and the decoder reconstructs the complete input. The reconstruction loss is calculated against the actual values to improve the model's accuracy in recovering missing information (Wardana et al. 2022)

In the air quality monitoring context, LSTMs are particularly effective for modeling temporal dependencies in data, such as predicting pollutant concentrations based on historical measurements. They can also handle missing data and noisy inputs, making them suitable for real-world applications.

LSTMs face some challenges, including the requirement for substantial labeled data for effective training and the risk of overfitting when the model is overly complex. They can also be computationally intensive, particularly for long sequences or deep architectures, and their learned representations are often difficult to interpret.

4.1.12 Transformers

Transformers are a type of deep learning architecture that has gained popularity in recent years, particularly in natural language processing tasks. They utilize self-attention mechanisms to capture relationships between input elements, allowing them to model long-range dependencies without relying on sequential processing.

Since the introduction of attention mechanisms in (Vaswani et al. 2017), there has been growing interest in transitioning LSTM-based approaches to transformer-based models for time series prediction, including air quality prediction. Attention mechanisms allow models to focus on relevant parts of the input data, which improves their ability to capture long-range dependencies without the limitations of sequential processing found in LSTMs. The original transformer architecture is presented in Fig. 6.

Transformers have allowed the development of new architectures that can outperform LSTMs in various tasks, including time series forecasting.

Transformers require large amounts of labeled data for training, the risk of overfitting if the model is too complex, and the difficulty in interpreting the learned representations. Additionally, transformers can be computationally expensive to train, especially for long sequences or deep architectures.

4.1.13 Ensemble and hybrid models

Ensemble and hybrid models combine multiple ML or DL techniques to improve prediction accuracy and robustness. These approaches integrate the strengths of different models while mitigating their individual weaknesses. Common ensemble methods include bagging, boosting, and stacking, while hybrid models may combine different architectures or integrate ML and DL techniques.

Ensemble and hybrid models are widely used in air quality monitoring to enhance the performance of multiple techniques, as will be discussed in the next sections.

4.2 Data imputation for handling missing values

In air quality monitoring, missing or erroneous data points may arise due to sensor failures, data transmission errors, or environmental disturbances. Properly treating these anomalies is essential to ensure the reliability of air quality assessments and subsequent analyses with the available data. The distribution of missing values in air quality datasets can follow complex patterns, as illustrated in Fig. 7, where black cells represent missing entries between different samples and variables.

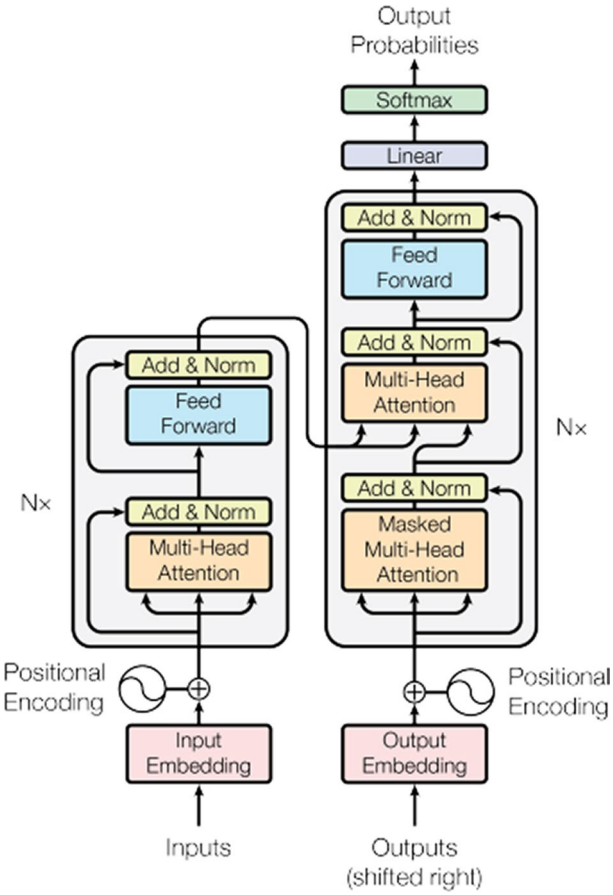


Fig. 6 The Transformer architecture, consisting of an encoder-decoder structure, as presented in (Vaswani et al. 2017). The encoder (left) processes input embeddings with positional encodings and passes them through multiple layers of multi-head attention and feed-forward networks with residual connections and layer normalization. The decoder (right) follows a similar structure but includes masked multi-head attention to prevent information leakage from future tokens. The final output probabilities are obtained through a linear layer followed by a softmax function

Studies have demonstrated that DL-based methods like autoencoders and denoising models (MIDIA, TCDA, IM-GAN) outperform traditional imputation in scenarios involving long or irregular gaps. Models that incorporate multivariate inputs or temporal context improve accuracy and reduce bias. New hybrid approaches such as BFMVI and logistic regression-based FTLRI further improve robustness and efficiency under high missing data rates. The study presented in (Shaadan and Rahim 2019) explored multiple imputation techniques, including linear interpolation, spline interpolation, exponential moving average, random replacement, Kalman filter, ARIMA, and mean before-after imputation. These methods have been widely used in air quality time series data to estimate missing values while ensuring the consistency of the dataset for further analysis. Each method presents distinct strengths, depending on the data distribution, the missingness pattern, and the analytical objectives. Although interpolation techniques are effective for small gaps in sequential data, Kalman filters and ARIMA models are more suitable for longer gaps and time-dependent trends.

Another research work, (Junninen et al. 2004), discussed a range of statistically based imputation methods, categorizing them into univariate, multivariate, hybrid, and multiple imputation strategies. Their findings emphasize that while univariate methods, such as

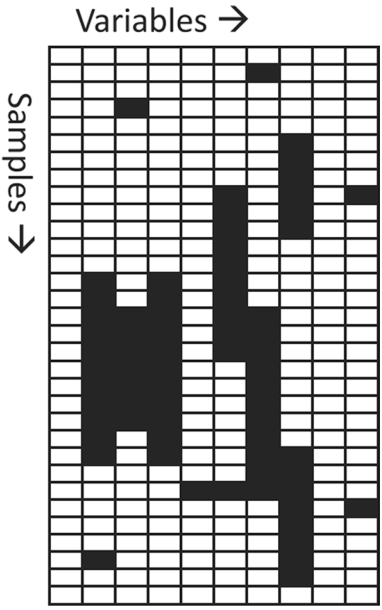


Fig. 7 Visualization of missing data patterns in air quality datasets. The black cells represent missing values across different samples and variables (Gómez-Carracedo et al. 2014)

linear and spline interpolation, effectively handle short gaps, their performance deteriorates with longer missing sequences. Multivariate techniques, in particular self-organizing maps (SOM) and multilayer perceptron (MLP), used statistical modeling to capture complex dependencies within air quality data, improving the accuracy of the imputation. The study also introduced a hybrid approach that applies interpolation for short gaps while using statistical learning models for longer gaps, balancing computational efficiency with accuracy. Their results also indicate that multiple imputation methods, which integrate the results of several statistical models, improve data consistency by accounting for estimation uncertainty.

Another widely used statistical approach involves imputation-based methods to handle missing values in time series air pollution data. The authors in (Junger and Leon 2015) proposed an Expectation-Maximization (EM) algorithm designed for multivariate time series, which integrates techniques such as ARIMA models, natural cubic splines, and regression-based filtering. Their study, conducted on PM_{10} concentration data Brazil, showed that EM-based imputation achieves lower RMSE and mean absolute error (MAE) values compared to simpler techniques, such as unconditional mean or median imputation, while preserving temporal dependencies. These results are evidence of the importance of statistical models in maintaining data integrity and minimizing bias in air quality assessments.

The study in (Gómez-Carracedo et al. 2014) evaluates statistical methods for handling missing data in air quality datasets, focusing on single and multiple imputation techniques applied to concentration measurements of pollutants such as PM_{10}, NO_2, and O_3. Single imputation methods, including mean imputation and regression-based approaches, provide computationally efficient solutions but often introduce bias, particularly when data are not missing at random. Multiple imputation techniques, such as multiple imputation by chained equations (MICE), expectation maximization, and Bayesian imputation, offer more robust statistical approaches by preserving data variability and reducing uncertainty. The results

show that while simpler methods can be effective for small gaps, more advanced statistical imputation techniques, particularly EM and Bayesian-based methods, achieve higher accuracy and reliability, making them suitable for air quality monitoring applications.

In contrast, advanced ML algorithms, such as KNN, DT, and DL models, can also learn the relationships between variables and provide more accurate imputation results (Alkabbani et al. 2022). For example, DL approaches, including autoencoders, have been used to reconstruct missing data by capturing complex nonlinear interactions in the dataset (Kim et al. 2021). These methods not only improve the completeness of the data set, but also enhance the performance of subsequent analyzes, such as the prediction of the pollutant concentration and the estimation of the air quality index (Wijesekara and Liyanage 2020).

The paper in (Agbo et al. 2022) also evaluates several ML-based imputation techniques to handle missing data in IoT sensor networks, focusing on environmental monitoring. The methods tested include EM imputation, KNN imputation, MissForest, regression imputation (RI), and a novel clustering-based approach called best fit missing value imputation (BFMVI). The study simulates different missing data rates (from 10% to 40%) and patterns, comparing the performance and computational complexity of each technique, with BFMVI showing the best accuracy but at a higher computational cost.

In air quality monitoring, DL effectively handles complex datasets, especially with missing data. Deep neural networks (DNN) can model intricate relationships within the data, outperforming traditional imputation methods. In (Wardana et al. 2022) authors present an autoencoder-based method is proposed to estimate missing values in air quality data. The model, which uses one-dimensional convolutional layers, is designed to capture both spatial and temporal behaviors of air pollutants by leveraging data from nearby monitoring stations. This approach does not require additional external features, such as weather or climate data, focusing solely on pollutant data. The study demonstrates that the proposed method significantly outperforms traditional univariate and multivariate imputation techniques, particularly in handling datasets with missing values of discontinuous and long intervals.

The study in (Samal et al. 2021) presents a DL framework that aims to predict $PM_{2.5}$ concentrations while managing missing data within a single training process. The proposed framework, called the temporal convolutional denoising autoencoder (TCDA), integrates temporal convolutional networks and denoising autoencoders (DAE) to handle multivariate sequential data and reconstruct missing values. The study conducts a comparative analysis of the effectiveness of the model in error reconstruction and prediction accuracy compared to other baseline models.

A different denoising autoencoder called missing data imputation denoising autoencoder (MIDIA) is presented in (Ma et al. 2020). MIDIA is a specialized DL model that aims to identify nonlinear relationships between missing and observed values, enhancing the accuracy of the imputation. The researchers introduced two methodologies derived from the MIDIA model: MIDIA-Sequential, which imputes missing values one attribute at a time, and MIDIA-Batch, which performs the imputation collectively. These strategies were compared with current imputation methods, showing good performance in managing various missing data patterns.

The work in (Wu et al. 2022) presents a novel approach to handling missing data in indoor air quality datasets. The proposed method, named inverse mapping generative adversarial network (IM-GAN), integrates several advanced ML components, including bi-directional recurrent neural networks (BRNN), DAEs, and GAN. The IM-GAN framework is designed

to capture both bidirectional temporal correlations and across-sensor correlations in multivariate time series data, while also effectively managing data distribution and redundancy issues. The study concludes that IM-GAN outperforms state-of-the-art methods in terms of accuracy of imputation on public indoor air quality (IAQ) datasets, addressing key technical challenges such as network saturation and approximation of data distribution.

Although state-of-the-art methods in missing data imputation for air quality monitoring often favor DL models, particularly autoencoders, there are other works that utilize traditional ML techniques. For example, the study in (Chen et al. 2022) presents a novel imputation approach for missing data in time series air quality data using logistic regression. Their method, named first five last three logistic regression imputation (FTLRI), focuses on capturing temporal correlations and attribute relationships to effectively fill in missing values. This approach is particularly designed for datasets with high missing rates and has shown notable improvements in terms of performance when compared to other traditional neural network-based imputation methods.

The handling of missing data in air quality monitoring remains a critical challenge, for which a wide range of statistical, ML, and DL approaches are available. Although statistical techniques offer simplicity and interpretability, ML and DL models provide more sophisticated solutions capable of capturing complex relationships within data. As air quality monitoring systems continue to evolve, the integration of multiple imputation strategies and the use of advances in AI can improve the reliability of the data. Overall, data imputation plays a foundational role in enabling downstream analysis and forecasting. The increasing use of hybrid and DL-based imputation methods reflects their growing importance. With missing data addressed, the next section explores how ML methods enhance sensor calibration and drift compensation.

4.3 Sensor calibration and drift compensation

Accurate sensor calibration is essential to ensure data accuracy in air quality monitoring. Over time, sensors can drift due to aging, environmental changes, or pollutant build-up. If not corrected periodically, this drift can can compromise data integrity and result in notable inaccuracies. ML techniques like multiple linear regression, RF, and ANN have been used for calibration. These techniques make it possible to automate and optimize calibration, providing real-time adjustments and compensating for sensor drift. These methods not only improve data accuracy and reliability but also optimize the performance of low-cost sensors, making them a viable option for large-scale environmental monitoring.

The study presented in (Zimmerman et al. 2018) investigates the calibration of low-cost air quality sensors, specifically the RAMP monitors, which measure pollutants such as CO, NO_2, O_3, and CO_2. The researchers compared three calibration approaches: a traditional laboratory-based univariate linear regression (LAB), an empirical multiple linear regression (MLR) that accounts for environmental variables, and a RF regression (RFR) ML model. The findings revealed that the RFR model achieved better results than both LAB and MLR, particularly in handling cross-sensitivities and environmental influences, resulting in more accurate and reliable pollutant measurements. The performance of the RFR model met the recommendations of the US Environmental Protection Agency (EPA) Air Sensor Guidebook, indicating its potential to improve low-cost sensor networks for air quality monitoring.

The authors in (Bush et al. 2022) also applied RFR as a calibration technique. This approach was used to refine the measurements of key pollutants, including NO_2, PM_{10}, and $PM_{2.5}$, which are often affected by varying environmental conditions. By implementing RFR, the study shows a substantial reduction in MAE, ranging from 37% to 94%, depending on the pollutant. The calibration achieved high levels of precision and the sensor accuracy improved to within ±2.6 ppb for NO_2, ±4.4 $\mu g/m^3$ for PM_{10}, and ±2.7 $\mu g/m^3$ for $PM_{2.5}$. These results show RFR's potential to improve low-cost sensor reliability and accuracy for large-scale monitoring.

Furthermore, the study in (Ali et al. 2021) applied ANN to calibrate low-cost sensor nodes with LoRaWAN IoT connectivity. ANN-based calibration showed improvements in reducing mean absolute percentage error (MAPE) and improving R^2 values for pollutants such as CO and NO_2, compared to traditional methods such as ordinarily least squares (OLS) and MLR. This shows the practical application of ML in managing sensor drift and environmental variability in low-cost air quality monitoring systems.

In (Park et al. 2021) the authors propose a new calibration model, HybridLSTM, which combines an LSTM with a DNN to address the limitations of conventional calibration methods for low-cost $PM_{2.5}$ sensors. The study compares the performance of the HybridLSTM model with benchmarks such as MLR and a standard DNN model. The results indicate that HybridLSTM demonstrates improved calibration accuracy, reducing RMSE by 41-60% relative to raw sensor data, 30-51% to MLR, and 8-40% to the DNN model.

The paper in (Tancev and Toro 2022) examines the use of variational Bayesian methods to improve the calibration of low-cost gas sensors. The study focuses on addressing challenges such as cross-sensitivities, environmental interference, and sensor aging. The authors implemented variational Bayesian Linear Regression (BLR) and Bayesian Neural Networks (BNN) to calibrate the sensors, incorporating uncertainty estimates into the calibration process. This approach helps identify when sensor measurements fall outside predefined confidence regions, indicating the need for recalibration or maintenance. The study suggests that Bayesian models can be a viable alternative to traditional calibration methods to improve the reliability of low-cost sensor networks.

Recent advances show that many calibration and drift correction techniques for low-cost air quality sensors use low-computation ML regression models, suitable for measuring stations' limited processing power. AI-based calibration has become a reliable way to extend the utility of low-cost sensors in diverse environments. Once calibrated, the systems must also detect and handle anomalous data, which is the focus of the next section.

4.4 Anomaly detection and outlier treatment

In air quality monitoring, anomalies and outliers can arise due to a number of factors, including sensor failures, environmental disturbances, or transient fluctuations in pollution levels. Detecting and managing these outliers is essential to maintaining the accuracy and reliability of data. ML techniques for anomaly detection help distinguish between genuine environmental anomalies and sensor errors, ensuring that downstream analyses such as pollution prediction and source attribution are based on accurate measurements. Advances in ML, DL, and traditional statistical techniques continue to improve anomaly detection and outlier management in air quality monitoring.

Statistical methods are widely used to detect outliers in air quality data due to their computational efficiency and effectiveness. As discussed in (Mahajan et al. 2020), techniques such as Z-score, interquartile range (IQR), Grubb's test, Hampel test, and Tietjen-Moore test have proven effective for incremental outlier detection. These methods are particularly valuable in resource-limited settings as they require fewer computational resources compared to ML models and can handle real-time data streams. This study suggests that statistical methods provide a fast and reliable alternative to more complex algorithms, making them suitable for low-cost air quality monitoring systems, especially in developing regions.

Another approach to the detection of outliers in air quality data involves functional data analysis (FDA). The study in (Martínez et al. 2014) applies this method to measurements of pollutants, including CO, NO_2, and SO_2, in the urban area of Langreo in northern Spain. It compares classical statistical methods, such as control charts and Box-Cox transformations, with the FDA model, which treats data as continuous time series rather than isolated points. The concept of functional depth is used to detect outliers by measuring the centrality of pollutant concentration curves. The results indicate that the FDA model is more effective in identifying real environmental anomalies, while classical methods often flag a higher number of outliers due to measurement errors or deviations that do not represent real pollution events.

Although statistical and functional data analysis methods offer efficient solutions for anomaly detection, ML techniques provide an adaptive approach capable of capturing more complex patterns in air quality data. Unlike statistical methods, which rely on predefined thresholds and assumptions about data distribution, ML models learn patterns from historical data and adapt to intricate relationships within the dataset, making them effective for detecting and predicting anomalies in air quality monitoring.

A hybrid model for AQI forecasting that integrates outlier detection, forecasting, and heuristic optimization is presented in (Wang et al. 2020). The Hampel identifier detects and corrects outliers in the data, while variational mode decomposition (VMD) decomposes the data into multiple components. These components are then predicted using an extreme learning machine (ELM) optimized by the sine cosine algorithm (SCA). The model's performance is evaluated using AQI datasets from Tianjin and Shenyang. The results show that the proposed hybrid model improves prediction accuracy compared to traditional models such as ARIMA and ELM, particularly in reducing MAPE and RMSE. The study presents the ability of the model to handle non-linear and non-stationary AQI time series data and emphasizes the benefits of combining outlier correction with ML-based forecasting methods.

Outlier detection and gap-filling techniques for low-cost air quality sensors are explored in (Ottosen and Kumar 2019), which applies ML algorithms over an 11-month period. The study evaluates two outlier detection methods: KNN for point outlier detection and ARIMA for contextual outlier detection. For gap filling, methods including linear interpolation, cubic spline fitting, and neural networks are applied. The results indicate that KNN and ARIMA effectively identify outliers, while linear interpolation outperforms other gap-filling techniques.

The study by (Hill and Minsker 2010) presents a data-driven approach to detect anomalies in streaming environmental sensor data using an autoregressive modeling technique. This method generates a prediction interval (PI) based on recent historical data and flags measurements that fall outside this range as anomalies. The approach is designed to work

incrementally, making it suitable for real-time applications where large volumes of data must be processed efficiently. The study evaluates multiple anomaly detection models, including multilayer perceptrons, single-layer linear networks, and nearest cluster predictors, and applies them to wind speed data from Corpus Christi, Texas. The results indicate that MLP-based anomaly detection, combined with a mitigation strategy that replaces detected anomalies with predicted values, yields appreciable improvements in data quality control.

Deep learning has emerged as a powerful approach for anomaly detection in air quality monitoring by capturing complex nonlinear relationships in large datasets. Autoencoders, widely recognized for their anomaly detection capabilities, excel in identifying outliers by learning compact representations of normal patterns and detecting deviations through reconstruction errors (Zhou and Paffenroth 2017). In particular, denoising autoencoders improve robustness by reconstructing clean signals from noisy inputs (Vincent et al. 2008), making them well suited for high-dimensional time series data and valuable for detecting sensor failures, transient fluctuations, and environmental anomalies in air quality monitoring systems.

A study presented in (Samal et al. 2021) proposed a TCDA network to improve air pollution prediction by addressing missing values and detecting outliers. The model integrates a temporal convolutional network with a DAE to extract features from complex air quality datasets while reconstructing missing or erroneous data. The study used a multivariate dataset from Beijing, China, covering pollutants such as $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3, along with meteorological variables. The TCDA model was evaluated against ML, DL, and statistical forecasting methods. The results indicate that TCDA achieved lower RMSE, MAE, and MAPE compared to other benchmark models, showing its effectiveness in handling outliers and reconstructing missing values.

In (Park et al. 2023) an approach based on DL is presented, with an unsupervised outlier detection framework for time-series indoor air quality data, using a long-short-term memory autoencoder. The autoencoder encoder compresses the input data into a low-dimensional latent space, extracting key features from the multivariate time series, while the decoder reconstructs the original data to detect reconstruction errors, which serve as indicators of potential outliers. Additionally, a subalgorithm that uses one-class support vector machines is used on the encoder-generated latent features, enabling enhanced refinement of the outlier detection procedure. The framework integrates these models using an ensemble method, which combines the outputs of both the LSTM-AE reconstruction error and the OC-SVM to enhance detection accuracy and minimize false positives. The approach was tested in real-world indoor air quality data, successfully improving anomaly detection while handling computational limitations.

The paper in (Wei et al. 2023) introduces a hybrid DL model combining LSTM networks and an autoencoder to address the challenge of detecting anomalies in indoor air quality data. The authors utilize LSTM cells to capture long-term dependencies within time series data and the autoencoder to determine an optimal anomaly threshold based on reconstruction loss. This approach is tested in the Dunedin CO_2 dataset, collected from New Zealand schools, which includes more than 247,000 CO_2 readings. The results indicate that the proposed model achieves a robust accuracy of 99.50%, outperforming other similar models. The study also examines the effectiveness of the LSTM-AE architecture in detecting anom-

alies using patterns in sequential data, showing that the model provides high precision and recall, particularly in detecting CO_2 anomalies in school environments.

A different LSTM autoencoder is described in (Rollo et al. 2023), which improves air quality monitoring by identifying and correcting anomalies in the raw data from low-cost sensors. The anomaly detection employs three approaches: sliding window anomaly detection, iterative data capture by the forgetting factor, and temperature and humidity-based detection. An LSTM autoencoder is used to confirm anomaly detection outcomes by understanding normal sensor behavior and detecting anomalies through reconstruction errors. Once anomalies are detected, a vector autoregression (VAR) model is utilized to recover the missing data. Deep learning is also applied in the calibration phase, where LSTM models forecast pollutant concentration levels from cleaned and repaired sensor data. This integrated method notably reduces errors and improves the accuracy of both anomaly detection and LSTM-based calibration models.

The work presented in (Jesus et al. 2021) explores the application of ML for reliable outlier detection in environmental monitoring systems, specifically under harsh and dynamic conditions. The authors introduce a methodology called ANN outlier detection (ANNODE), which leverages ANN to model sensor behavior, exploiting correlations between multiple sensors measuring related environmental parameters. The method detects sensor faults in real time by comparing observed measurements to predicted values, allowing the system to distinguish between true sensor failures and anomalies caused by natural environmental events. The study evaluates the methodology using real-world data sets from an aquatic monitoring system that measures temperature and salinity. The results show that ANNODE not only detects outliers with high accuracy but also improves data quality by correcting erroneous measurements. Compared with other state-of-the-art outlier detection techniques, ANNODE outperforms in terms of accuracy and dependability, particularly in challenging environments.

The reviewed articles present innovative methods for identifying anomalies and managing outliers in air quality monitoring. Techniques such as autoencoders, LSTMs, and hybrid models that integrate statistical techniques with ML stand out for their effectiveness in anomaly detection and data gap interpolation. Moreover, basic statistical methods offer a balance between computational efficiency and accuracy in low-cost and resource-limited settings. Effective anomaly detection supports reliable forecasting and risk assessment. The next section demonstrates how these enhanced data streams are applied in air quality index estimation.

4.5 Air quality index (AQI) estimation

The AQI is a numerical scale used by government agencies and health organizations worldwide to communicate how polluted or clean the air is. It translates complex data on different pollutants into a simple, standardized number, helping the public quickly understand current air quality conditions and associated health risks.

Accurately estimating the AQI is essential for effective air quality monitoring, as it provides a standardized measure to inform the public about potential health risks. AQI is typically calculated by weighting the concentration of pollutants, including $PM_{2.5}$, PM_{10}, carbon monoxide (CO), nitrogen dioxide (NO_2), and ozone (O_3), reflecting the impact of exposure to these pollutants on health. ML, especially regression-based techniques, has proven valu-

able in predicting pollutant levels and estimating AQI from historical and real-time data. These models not only improve the accuracy of AQI estimation but also facilitate real-time forecasting, enhancing decision-making in environmental management.

A notable example of ML-based AQI estimation is found in (Kok et al. 2017), where AQI is calculated by mapping pollutant concentrations O_3 and NO_2 to AQI values based on threshold ranges established by the US EPA. These thresholds classify air quality into categories such as “Good", “Moderate", “Unhealthy for Sensitive Groups", “Unhealthy", “Very Unhealthy", and “Hazardous." The general AQI for a given period is determined by selecting the highest AQI value among the measured pollutants, representing the most severe air quality condition. In this study, an LSTM network is used to predict pollutant concentrations, which are then used to estimate the AQI. The model assigns alarm levels based on AQI thresholds, with color-coded alerts for ease of interpretation. The precision of AQI estimation is evaluated using metrics such as precision, recall, and F1 score, providing information on the model’s ability to classify air quality accurately.

Based on the use of LSTM networks, (Janarthanan et al. 2021) explores a hybrid model that combines support vector regression (SVR) with LSTM networks to predict AQI. This model uses pollutant concentrations including $PM_{2.5}$, NO_2, SO_2, CO, and O_3, but also incorporates a feature extraction process using the gray-level co-occurrence matrix (GLCM) method to optimize input data. While the LSTM network captures temporal dependencies in pollutant data, SVR is utilized to handle nonlinear relationships in concentration levels. The hybrid approach improves the prediction accuracy of the AQI compared to traditional models, achieving lower Mean Squared Error (MSE) and higher R^2 values.

Expanding existing methods for AQI prediction, (Hossain et al. 2021) presents a hybrid approach that merges CNN with LSTM networks. This model addresses both spatial and temporal dependencies in pollutant data by using CNN layers to extract spatial features and LSTM layers to model temporal relationships in time series data. Pollutants such as $PM_{2.5}$, NO_2, and SO_2 are analyzed to predict future AQI values more effectively by considering both historical patterns and time-dependent trends. This hybrid CNN-LSTM model surpasses traditional methods such as linear regression and ARIMA models, achieving lower MSE and higher R^2 values across different pollutants. As a result, it has been shown to be a highly reliable tool for real-time AQI prediction, particularly suited to the needs of urban environments and smart cities.

The study in (Kim et al. 2022) uses LSTM networks to capture temporal dependencies in pollutant concentration data, using LSTM’s ability to model long-term relationships in time series data for pollutants such as PM_{10}, $PM_{2.5}$, and NO_2. Following the LSTM-based pollutant level predictions, a DNN is used to predict the AQI, integrating the predicted pollutant concentrations to generate the final AQI score. The DNN is designed to handle the non-linear relationships between multiple pollutants and their combined effect on air quality. The DL elements of the model, namely LSTM and DNN, jointly provide accurate air quality predictions in temporal and spatial dimensions. This could be verified through performance metrics including correlation coefficient (CC) and normalized root mean square error (NRMSE).

In both studies (Liu et al. 2019) and (Maltare and Vahora 2023), ML techniques are used to predict air quality, although the approaches differ in complexity and methodology. In (Liu et al. 2019), the authors rely on traditional ML models, specifically SVR and RFR. These models, although simpler, are highly effective and are applied to predict AQI in Beijing

using SVR, which captures the complex relationships between pollutants and AQI, and to forecast NOx concentrations in an Italian city using RFR, which mitigates overfitting by building and averaging decision trees. The study provides evidence that these methods can generate accurate predictions without the computational demands of DL, with performance assessed using RMSE and R^2 metrics. Based on this, the (Maltare and Vahora 2023) study presents a more complex hybrid approach that integrates multiple ML models including the support vector machine with a kernel of radial basis function, seasonal autoregressive integrated moving average (SARIMA), and LSTM to predict the AQI in Ahmedabad, India. This approach covers a dataset from 2015 to 2021, including key pollutants such as $PM_{2.5}$, PM_{10}, CO, and NO_2. By combining the ability of SVM to handle non-linear relationships, the strength of LSTM to capture temporal dependencies, and the utility of SARIMA to identify seasonal trends, the hybrid model offers more robust predictions. However, SARIMA performed poorly in predicting short-term air quality indices compared to SVM and LSTM, and model performance was evaluated using RMSE, R^2, and MSE. In general, these studies summarize the landscape of air quality prediction, from traditional models to hybrid approaches, emphasizing the trade-offs between simplicity and accuracy.

Most AQI estimation techniques use ML or DL methods. Traditional ML models like SVR and RFR efficiently handle nonlinear pollutant data relationships. In contrast, DL techniques such as LSTM and CNN capture complex temporal and spatial dependencies. Hybrid ML and DL approaches enhance AQI predictions, especially in dynamic, large-scale urban environments. Building upon these estimation techniques, short-term forecasting takes the next step by projecting air quality conditions into the near future. This capability is critical for enabling early warnings and real-time decision-making in public health and environmental management. The following section explores the models and strategies used to forecast air pollution levels with high temporal resolution.

4.6 Short-term Air quality forecasting

Short-term air quality forecasting is important for real-time monitoring and mitigation of pollution, especially in urban areas where pollutant concentrations can fluctuate rapidly. Time-series analysis based on historical data allows identification of trends, seasonality, and irregular variations in pollutant levels. In addition, these methods facilitate the detection of outliers, the imputation of missing data, and the identification of anomalies, contributing to more accurate datasets. Models such as ARIMA, LSTM networks, and other ML techniques have consistently demonstrated their potential to improve prediction accuracy and data reliability.

A notable example of this is the study by (Du et al. 2021), which introduced a hybrid DL model that combines one-dimensional CNN and bidirectional long short-term memory (Bi-LSTM) networks to forecast $PM_{2.5}$ levels. This hybrid approach, which leverages spatial and temporal dependencies in air quality data, successfully addresses the nonlinear and dynamic characteristics of environmental data. When tested in $PM_{2.5}$ and Beijing urban air quality datasets, the hybrid model substantially outperformed traditional ML techniques such as SVR and ARIMA, as well as other models such as LSTM and GRU, confirming its reliability in forecasting air quality over different periods.

Similarly, the (Freeman et al. 2018) study focused on predicting 8-hour-averaged O_3 concentrations using LSTM networks, showing the importance of temporal dependencies

in air quality data. The model also incorporated novel data imputation techniques to handle missing values and used decision trees to reduce the feature set from 25 to 5, improving efficiency. With the ability to forecast O_3 concentrations up to 72 h in advance, the LSTM model achieved MAE values below 2 ppb, outperforming ARIMA and forward propagation neural networks (FFNN), especially in handling long-term dependencies and nonlinear patterns.

Building on these advances, (Li et al. 2016) introduced a spatio-temporal deep learning (STDL) model for $PM_{2.5}$ prediction. The STDL model, based on a stacked autoencoder (SAE) architecture, was designed to automatically extract spatial and temporal features from data collected at multiple monitoring stations. Compared with traditional models such as spatio-temporal artificial neural networks (STANN), ARMA, and SVR, the STDL model yielded superior results by effectively capturing complex spatiotemporal dependencies, leading to more accurate forecasts across different stations and locations.

To further improve LSTM-based prediction, (Navares and Aznarte 2020) employed LSTM networks to predict pollutants such as CO, NO_2, O_3, PM_{10}, SO_2, and airborne pollen in Madrid. This study was differentiated using multiple LSTM configurations, fully connected LSTM, pooled pollutant LSTM, and single-pool pollutant LSTM, to model spatial and temporal patterns at multiple monitoring stations simultaneously. Statistical tests showed that LSTM-based configurations reduced prediction errors, outperforming traditional methods such as RFR and linear regression, particularly for day-ahead forecasts.

The multitask learning framework presented in (Xu and Yoneda 2021) extends the application of LSTM by integrating an LSTM autoencoder model to predict $PM_{2.5}$ levels at multiple monitoring stations in Beijing. The multitask approach allows the model to share learning across stations, capturing underlying relationships at different locations. By encoding meteorological data via a SAE, the model reduces errors (RMSE, MAE) in both single-step and multi-step forecasting tasks, particularly during periods of pronounced variability in air quality. The multitask model also demonstrated computational efficiency, reducing processing time by nearly an order of magnitude compared to models that addressed each task independently.

The authors in (Xayasouk et al. 2020) presented a comparative study between LSTM networks and deep autoencoders to predict PM_{10} and $PM_{2.5}$ concentrations in Seoul. The study presented the strengths of LSTM in capturing temporal dependencies, producing lower RMSE values of 11,113 for PM_{10} and 12,174 for $PM_{2.5}$, compared to the DAE model, which achieved RMSE values of 15,038 and 15,437, respectively. Although the LSTM model showed superior performance, particularly in terms of accuracy, the DAE model provided a computationally cheaper alternative. The authors suggest further improvements by integrating GIS-based spatial data for more accurate predictions.

The paper published in (Liang et al. 2023) introduces AirFormer, a transformer-based model specifically designed to predict air quality in China at fine spatial granularity, using data from thousands of locations. The model takes a two-stage approach. In the first stage, a bottom-up deterministic process captures spatial and temporal dependencies using multi-head self-attention mechanisms. This allows the model to focus on the relationships between geographically dispersed monitoring stations via dartboard spatial MSA (DS-MSA) for spatial relationships and causal temporal MSA (CT-MSA) for temporal dependencies. In the second stage, a top-down stochastic process with latent variables is used to manage the inherent uncertainty in air quality data, providing a more accurate and adaptive predic-

tion framework. By integrating air quality and meteorological data, AirFormer effectively models the complexities of pollutant levels in different regions, and its design is optimized to handle large-scale datasets efficiently. Comparative experiments show that AirFormer consistently outperforms traditional models, such as DeepAir and $PM_{2.5}$-GNN, and reduces prediction errors by 5%-8% in 72-hour forecasts. The model performs particularly well under stable conditions and sudden fluctuations in air quality, as tested on a four-year dataset collected from 1,085 monitoring stations in mainland China. The two-stage framework not only captures spatial and temporal dynamics but also enhances robustness in handling uncertainty, making AirFormer a marked advancement over existing approaches for large-scale air quality prediction.

The study conducted by (Zhang and Zhang 2023) introduces a novel DL model using limited attention-based transformer networks (STN) to improve air quality prediction, specifically focusing on $PM_{2.5}$ levels. The model is designed to handle the limitations of traditional methods, such as the inability to model long-term dependencies in time-series data. The STN model consists of encoder and decoder layers and uses a multi-head sparse attention mechanism to efficiently learn long-term dependencies while reducing computational complexity. Extensive experiments were conducted on two datasets from Beijing and Taizhou. The findings indicate that the proposed model surpasses traditional techniques like ARIMA, SVR, as well as more recent deep learning methods such as LSTM and CNN. The model achieved superior results in both short- and long-term prediction, particularly in scenarios involving multi-step predictions for the next 48 h. The sparse attention mechanism allowed the model to reduce time complexity, making it a more efficient solution for large-scale air quality prediction tasks.

In contrast, the authors in (Wang et al. 2022) explore the potential of using mobile devices for air quality monitoring by proposing a transformer-based solution named the dual output vision transformer (DOViT). This approach employs a multihead self-attention (MSA) mechanism to automatically extract relevant features from images captured by mobile devices, enabling simultaneous prediction of both the AQI level and its numerical value. This approach improves previous CNN-based methods by improving classification accuracy, while offering greater flexibility and frequency of use in air quality monitoring. The model, trained on a dataset collected by mobile devices, obtained improved accuracy compared to traditional CNN architectures such as AlexNet and ResNet, and its efficiency enables it to be a viable tool for real-time air quality monitoring. The study shows the potential of using mobile devices as portable air quality monitoring stations, which offer a more accessible and dynamic method of environmental protection.

The paper published in (Chen et al. 2022) presents a hybrid CNN-Transformer model aimed at improving the prediction of ozone concentrations by capturing complex, non-linear relationships between ozone and various environmental factors such as NOx, SO_2, CO, and meteorological data. The model combines CNN's local feature extraction capability with the long-term dependency capturing strengths of the Transformer architectures. CNN focuses on learning local dependencies in the data, while Transformer applies attention mechanisms to capture global patterns, making the model more effective in handling multivariate time series. The study uses data from 14 monitoring stations in Beijing that span 2014 to 2021. Experimental results indicate that the hybrid model consistently outperforms traditional models such as LSTM, CNN-LSTM, and standalone Transformer architectures in short- and long-term ozone prediction tasks, particularly when handling complex multi-

variate datasets. The model architecture allows a more complete understanding of the interactions between pollutants and meteorological factors, leading to more accurate predictions.

Another paper published in (Li et al. 2022) proposes a DL model called TSF-Transformer, designed for time series forecasting of heavy-duty truck exhaust emissions. Unlike traditional methods, it focuses on predicting temperature and pressure changes in tailpipes using real-time sensor data from approximately 12,000 trucks. TSF-Transformer uses a multi-head self-attention mechanism to process data in parallel, offering advantages over models such as LSTM, which are limited by sequential processing. This transformer-based model encodes input data and uses an encoder-decoder structure for prediction, with the aim of improving both prediction speed and accuracy. Additionally, it visualizes the predictions through heat maps, allowing a more intuitive monitoring of the emissions over time. The paper compares TSF-Transformer with existing models such as SVM, XGBoost, LSTM and Autoformer, demonstrating superior performance in terms of emissions prediction for heavy trucks, particularly in terms of computational efficiency and speed.

The authors in (Cui et al. 2023) presented a study focused on predicting atmospheric $PM_{2.5}$ concentrations by comparing two DL techniques: a transformer model and a CNN-LSTM attention mode l. The study aims to develop more accurate predictive models for air quality, using historical data from 12 monitoring stations in Beijing. Both models incorporate meteorological factors such as temperature, wind, and humidity, along with pollution data. The transformer model was modified by retaining only the encoder part, combining time series with meteorological and pollutant features, and integrating a multihead attention mechanism to improve its ability to capture long-range dependencies. The findings indicated that the transformer model performs better than the CNN-LSTM attention model, especially in situations involving sudden meteorological changes. This makes it more appropriate for both seasonal and long-term forecasting. This model showed improved accuracy and robustness, especially during seasons with complex pollution trends such as fall and winter, indicating its potential to improve air quality forecasting.

The short-term air quality prediction is progressing due to the integration of advanced DL methodologies. Techniques such as LSTM and transformer-based models have been recognized for substantially refining the prediction of pollutant levels by exploiting spatial and temporal trends. These techniques, when coupled with attention mechanisms, offer superior precision in managing complex and nonlinear environmental data sets, and they present scalable solutions for wide-ranging monitoring systems.

5 Discussion

5.1 Hardware for air quality monitoring

Section 3 of this paper explores the hardware platforms, sensors, and communication protocols integral to air quality monitoring systems. Many of the reviewed studies developed low-cost prototypes, often based on microcontrollers or single-board computers. These systems range from simple setups, such as components mounted on a single-board computer such as Fig. 8, to more sophisticated designs featuring custom-built enclosures (such as Fig. 9) and fully integrated system-on-a-board PCB solutions like Fig. 10.

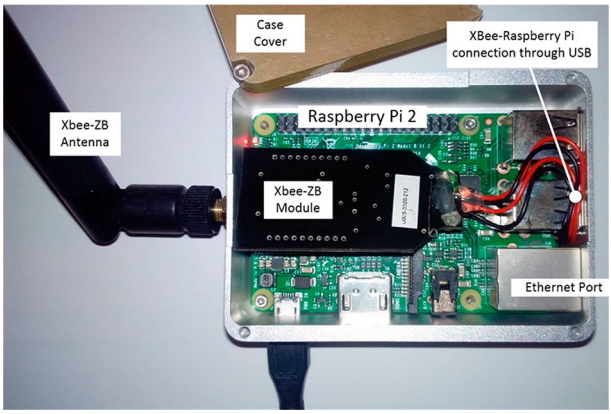
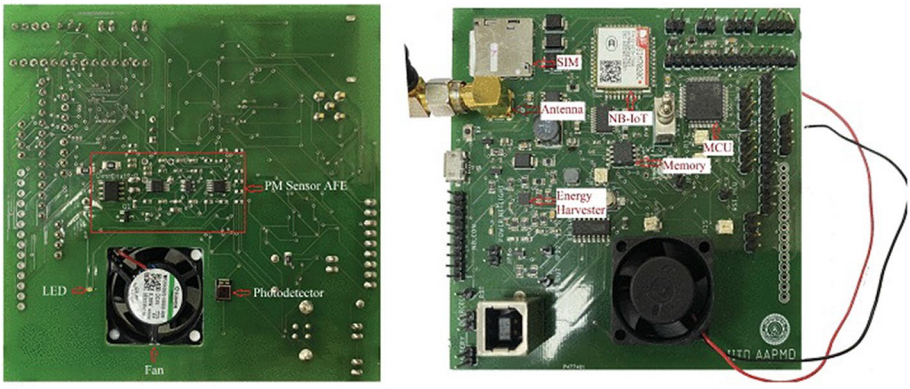


Fig. 8 A picture of an XBee gateway built around a Raspberry Pi 2, presented by Benammar et al. This device acts as a bridge between the XBee sensors used for air quality monitoring and the cloud, where the data is sent to a remote web server (Benammar et al. 2018)

Fig. 9 This DIY air quality monitoring system, presented by Collado et al., exemplifies the use of fast prototyping techniques such as 3D printing and CNC PCB fabrication. Designed as a monolithic station, this custom-built module was designed, built and deployed in an outdoor environment for air quality monitoring (Collado et al. 2024)



(a) Bottom side of the board (b) Top side of the board

Fig. 10 This PCB prototype, developed by Das et al. and named the Air Pollution Monitoring Device (APMD), exemplifies a custom-built system designed to meet specific project requirements. The APMD integrates all essential components into a single board, including a microcontroller, wireless communication modules, power management, and other elements, effectively functioning as a system-on-a-board (Das et al. 2022)

These prototypes face challenges such as power consumption, data transmission, and sensor calibration, all of which are influenced by hardware choices. Table 1 summarizes the air quality monitoring systems published between 2016 and 2024, identifying trends in technology and design. Research activity peaked in 2019 and 2022, reflecting an increased interest in low-cost, scalable solutions. Most of the systems targeted indoor or outdoor use, with only one study addressing both (Palomeque-Mangut et al. 2022).

As can be seen in Table 1, researchers often select sensors and protocols based on specific project requirements. In the microcontroller domain, designs frequently utilize Atmel microcontrollers and ESP modules, which are favored for their efficiency and cost-effectiveness compared to single-board computers like Raspberry Pi. While microcontrollers are more affordable and energy-efficient, single-board computers offer greater processing power and memory, making them suitable for more demanding applications.

Table 2 compares the most commonly used microcontrollers and single-board computers in air quality monitoring systems, detailing their processing power, memory, storage, and compatibility with programming languages and operating systems. These devices can be grouped into two categories based on their computing resources. Microcontrollers typically feature flash memory sizes ranging from 32 KB to 1 MB, RAM from 2 KB to 128 KB, and CPU frequencies between 16 MHz and 80 MHz. In contrast, single-board computers offer significantly higher resources, with flash memory dependent on SD card size, RAM ranging from 64 MB to 1 GB, and CPU frequencies from 400 MHz to 1.2 GHz.

Single-board computers can run full operating systems like Linux, whereas microcontrollers are typically limited to lightweight real-time operating systems (RTOS) or may operate without an OS entirely. While some microcontrollers support high-level programming languages like MicroPython, most rely on low-level languages such as C or C++. These constraints play an important role in the design of air quality monitoring systems, as they directly affect the complexity and flexibility of software for data processing and analysis.

Hardware capabilities significantly influence the choice of algorithms and models that can be implemented, particularly when considering the deployment of AI techniques on edge devices, as discussed later in this section.

Table1 lists the communication protocols reported in IoT-based air quality monitoring systems, highlighting their presence across the reviewed studies. While WiFi (IEEE 802.11) appears as the most commonly adopted protocol due to its widespread availability and ease of integration, other protocols such as LoRa, NB-IoT, and GPRS-designed for long-range and low-power communication-are mentioned less frequently. A more detailed comparative analysis of these protocols, particularly in terms of their technical characteristics and suitability for IoT applications, is provided in Table 3.

As shown in Table 3, the choice of communication protocol depends on factors such as range, power consumption, security requirements, and data transfer rates. Each protocol offers unique advantages and limitations, making it crucial to select the one that is best suited for the specific application. Communication protocols also have an impact on the decision of whether to use edge or cloud processing, as discussed later in this section.

5.2 Artificial intelligence in air quality monitoring

AI techniques play a pivotal role in air quality monitoring, addressing challenges such as missing data, sensor calibration, anomaly detection, air quality index estimation, and

Table 1 Summary of air quality monitoring systems presented in various publications over the past 8 years

Categories	Studies
Publication year	2016: (Zheng et al. 2016)
	2017: (Kumar and Jasuja 2017)
	2018: (Benammar et al. 2018), (Martín-Garín et al. 2018)
	2019: (Dhingra et al. 2019), (Husein et al. 2019), (Pradityo and Surantha 2019), (Sai et al. 2019), (Zhao et al. 2019), (Marques and Pitarma 2019)
	2022: (Das et al. 2022), (Jabbar et al. 2022), (Purbakawaca et al. 2022), (Palomeque-Mangut et al. 2022)
	2024: (Collado et al. 2024)
Indoors/outdoors	Indoors: (Benammar et al. 2018; Martín-Garín et al. 2018; Pradityo and Surantha 2019; Sai et al. 2019; Zhao et al. 2019; Marques and Pitarma 2019)
	Outdoors: (Zheng et al. 2016; Kumar and Jasuja 2017; Dhingra et al. 2019; Husein et al. 2019; Das et al. 2022; Jabbar et al. 2022; Purbakawaca et al. 2022; Collado et al. 2024)
	Indoor/outdoors: (Palomeque-Mangut et al. 2022)
Microcontroller or microprocessor	ATmega1281: (Benammar et al. 2018)
	ATmega2560: (Purbakawaca et al. 2022)
	ATmega32U4 + Atheros AR9331: (Collado et al. 2024)
	ATmega328P: (Kumar and Jasuja 2017; Dhingra et al. 2019; Husein et al. 2019; Pradityo and Surantha 2019; Jabbar et al. 2022; Sai et al. 2019)
	ESP-01: (Sai et al. 2019)
	ESP-8266: (Martín-Garín et al. 2018; Sai et al. 2019; Marques and Pitarma 2019; Das et al. 2022; Purbakawaca et al. 2022)
	PIC32MM0256GPM048: (Palomeque-Mangut et al. 2022)
	Raspberry Pi 2: (Kumar and Jasuja 2017)
	Raspberry Pi 3: (Pradityo and Surantha 2019)
	STM32F103C8T6: (Zheng et al. 2016), (Zhao et al. 2019)
	STM32L476RG: (Das et al. 2022)
Communication protocols	BLE + Smartphone: (Palomeque-Mangut et al. 2022)
	Ethernet (IEEE 802.3): (Collado et al. 2024)
	GPRS: (Zhao et al. 2019), (Purbakawaca et al. 2022)
	GSM: (Purbakawaca et al. 2022)
	LoRa: (Husein et al. 2019), (Zhao et al. 2019)
	LoRaWAN: (Jabbar et al. 2022)
	MQTT: (Kumar and Jasuja 2017)
	NB-IoT: (Zhao et al. 2019), (Das et al. 2022)
	RF (IEEE 802.15.4k): (Zheng et al. 2016)
	RS485: (Zhao et al. 2019)
	WiFi (IEEE 802.11): (Kumar and Jasuja 2017; Martín-Garín et al. 2018; Dhingra et al. 2019; Sai et al. 2019; Marques and Pitarma 2019; Zhao et al. 2019; Das et al. 2022; Purbakawaca et al. 2022; Collado et al. 2024)
	Xbee (IEEE 802.15.4): (Benammar et al. 2018)

Table 1 (continued)

Categories	Studies
Sensors	Alphasense AFE-A4 (NO_2, SO_2, CO, O_3): (Das et al. 2022)
	Anemometer (Wind Speed): (Purbakawaca et al. 2022)
	BME280 (Temperature, Humidity, Barometric pressure): (Benammar et al. 2018), (Martín-Garín et al. 2018)
	BME680 (Temperature, Humidity, Pressure, VOCs): (Palomeque-Mangut et al. 2022; Collado et al. 2024)
	BMP180 (Barometric pressure): (Kumar and Jasuja 2017), (Martín-Garín et al. 2018)
	CCS811 (CO_2, VOCs): (Palomeque-Mangut et al. 2022)
	DHT11 (Temperature, Humidity): (Das et al. 2022), (Jabbar et al. 2022)
	DHT22 (Temperature, Humidity): (Kumar and Jasuja 2017; Martín-Garín et al. 2018; Collado et al. 2024)
	GP2Y1010AU0F (PM_{10}): (Praditya and Surantha 2019)
	HPMA115S0-XXX ($PM_{2.5}$, PM_{10}): (Palomeque-Mangut et al. 2022)
	iAQ-Core (CO_2, VOCs): (Palomeque-Mangut et al. 2022)
	MH-Z19 (CO_2): (Martín-Garín et al. 2018)
	MICS-4514 (CO, NO_2): (Jabbar et al. 2022)
	MICS-6814 (CO, NO_2, C_2H_5OH, H_2, NH_3, CH_4, C_3H_8, C_4H_{10}): (Marques and Pitarma 2019; Collado et al. 2024)
	MPL-3115A2 (Pressure): (Purbakawaca et al. 2022)
	MQ131 (O_3): (Collado et al. 2024)
	MQ135 (CO_2, Ammonia, Smoke): (Kumar and Jasuja 2017; Dhingra et al. 2019; Husein et al. 2019; Praditya and Surantha 2019; Jabbar et al. 2022; Sai et al. 2019)
	MQ136 (SO_2): (Jabbar et al. 2022; Collado et al. 2024)
	MQ7 (CO): (Dhingra et al. 2019), (Husein et al. 2019), (Sai et al. 2019), (Purbakawaca et al. 2022)
	MQ9 (CO): (Kumar and Jasuja 2017), (Dhingra et al. 2019), (Jabbar et al. 2022), (Sai et al. 2019)
	OPC N3 (CO): (Das et al. 2022)
	PMS3003 ($PM_{2.5}$, PM_{10}): (Jabbar et al. 2022)
	PMS5003 ($PM_{2.5}$, PM_{10}): (Zhao et al. 2019)
	PMS5005 ($PM_{2.5}$, PM_{10}): (Zheng et al. 2016)
	Rain Gauge (Rainfall): (Purbakawaca et al. 2022)
	SDS011 ($PM_{2.5}$, PM_{10}): (Purbakawaca et al. 2022; Collado et al. 2024)
	SGP30 (CO_2, VOCs): (Palomeque-Mangut et al. 2022)
	SHT20 (Temperature, Humidity): (Zheng et al. 2016)
	SHT21 (Temperature, Humidity): (Martín-Garín et al. 2018)
	SHT30 (Temperature, Humidity): (Zhao et al. 2019)
	Si7021 (Temperature, Humidity): (Purbakawaca et al. 2022)
	Wind vane (Wind Direction): (Purbakawaca et al. 2022)
	ZMOD4410 (VOCs): (Palomeque-Mangut et al. 2022)

short-term forecasting. Table 4 summarizes the AI methods used, ranging from statistical approaches to machine learning and deep learning techniques.

Machine learning methods, such as DT and RF, are widely used for classification and regression tasks. Deep learning models, including CNN and LSTM networks, excel in han-

Table 2 Comparison of microcontrollers and single-board computers used for air quality monitoring applications: this table presents the technical specifications of various computing platforms that have been used for the development of air quality monitoring systems

Devices	Brand	Type	Flash memory	RAM	Frequency	OS capability	High-level language support
ATmega1281 (Microchip Technology Inc. 2016a)	Atmel/Microchip	Microcontroller (8-bit AVR)	128 KB	8 KB	16 MHz	No	No
ATmega2560 (Microchip Technology Inc. 2016b)	Atmel/Microchip	Microcontroller (8-bit AVR)	256 KB	8 KB	16 MHz	No	No
ATmega32U4 + AR9331 (Arduino Inc. 2020)	Atmel + Atheros	8-bit AVR MCU + MIPS SoC	32 KB (MCU) / 16 MB & SD Card	2.5 KB / 64 MB	16 / 400 MHz	Yes (Linux on AR9331)	Yes (Linux side)
ATmega328P (Microchip Technology Inc. 2016a)	Atmel/Microchip	Microcontroller (8-bit AVR)	32 KB	2 KB	16 MHz	No	No
ESP-01 (AI-Thinker 2020)	Espressif	Microcontroller (ESP8266 SoC)	512 KB	80 KB	80 MHz	No	Yes (MicroPython, NodeMCU, Lua)
ESP-8266 (Espressif Systems 2020)	Espressif	Microcontroller (ESP8266 SoC)	512 KB - 1 MB	80 KB	80 MHz	No	Yes (MicroPython, NodeMCU, Lua)
PIC32MM0256GPM048 (Microchip Technology Inc. 2016c)	Microchip	Microcontroller (32-bit PIC32MM)	256 KB	32 KB	64 MHz	No	No
Raspberry Pi 2 (Raspberry Pi Foundation 2015)	Raspberry Pi Foundation	SBC (ARM Cortex-A7)	SD card	1 GB	900 MHz	Yes (Linux)	Yes
Raspberry Pi 3 (Raspberry Pi Foundation 2016)	Raspberry Pi Foundation	SBC (ARM Cortex-A53)	SD card	1 GB	1.2 GHz	Yes (Linux)	Yes
STM32F103C8T6 (STMicroelectronics 2021a)	ST Microelectronics	Microcontroller (ARM Cortex-M3)	64 KB	20 KB	72 MHz	No (RTOS possible)	Yes (MicroPython)
STM32L476RG (STMicroelectronics 2021b)	ST Microelectronics	Microcontroller (ARM Cortex-M4)	1 MB	128 KB	80 MHz	No (RTOS possible)	Yes (MicroPython)

It describes hardware characteristics such as processing capability, memory capacity, and operating system support across different microcontrollers and single-board computers

dling complex data but require more computational resources. Hybrid approaches that combine multiple techniques often deliver superior performance.

5.2.1 Data imputation

Data imputation addresses missing values in datasets, which may occur due to sensor failures, communication issues, or other factors. Specialized techniques have been developed for this task beyond traditional statistical and AI methods (see Table 4). Methods such as EM and MICE are considered data imputation algorithms and are widely applied across various fields.

Table 5 summarizes the reviewed papers based on key aspects, including target variable, support variables, location, timespan, evaluation metrics, sampling rate, and missing rates. These aspects are fundamental because they influence the approach followed in each study. The target variable, usually a pollutant with missing data, is the most critical element, as it determines the type of data and the applicable imputation methods. It may involve a single pollutant or a combination of pollutants.

Support variables, such as additional pollutants or meteorological data, help determine whether a univariate or multivariate approach is used. An univariate imputation method focuses solely on the target variable, while a multivariate approach considers both the target and support variables. The choice of method depends on the availability of data and the relationships between variables.

The location and time span provide context for the data collection conditions. Performance is primarily assessed using error metrics like RMSE and MAE, with the problem treated as a regression task due to the prediction of continuous values. The missing rate, representing the percentage of absent data, is crucial because it affects the ability of the technique to reconstruct the dataset.

The table shows that studies vary in target variables, evaluation metrics, and missing rates. RMSE and MAE are the most common metrics, while missing rates of 5%, 10%, 20%, and 25% are frequently applied. In some cases, cleaned datasets were used with artificially introduced missing values at these rates, whereas in other studies naturally occurring missing values were addressed.

Table 6 presents the results of missing data imputation, where bold numbers denote the missing rates employed by each group of research studies. The most common target variables were $PM_{2.5}$ and PM_{10}. Most studies focused on imputing a single pollutant, while a few examined multiple pollutants and reported separate metrics for each. Although AI techniques have gained popularity in air quality monitoring, traditional statistical methods and established imputation algorithms continue to deliver superior results. Among AI methods, autoencoders emerged as the most promising approach, proving to be the most effective method in 4 out of 11 studies.

Comparisons across studies are challenging due to the diversity of target variables and the differing magnitudes of the data, which complicates the direct selection of the best imputation method since RMSE and MAE values are influenced by the scale of the target variable. However, the results indicate that AI-based techniques tend to outperform traditional statistical methods in most cases, with 7 of 11 studies reporting better results. Additional research is needed in which each method is applied to the same dataset under identical conditions to draw a definitive conclusion.

Table 3 Comparison of IoT communication protocols: advantages and limitations of commonly used wireless communication protocols for IoT in air quality monitoring

Protocol	Advantages	Limitations	References
Radio Frequency Communication	Versatile with various frequency bands available Can offer long range connectivity in sub GHz bands Often low cost and simple to implement	Susceptible to interference Performance highly dependent on environmental factors May require licensing in some bands	(Ganesh and Venkataraman 2021)
Bluetooth	Widely adopted and supported by many devices Low power consumption Easy pairing for short range communications	Limited range (typically up to 10 ms) Not ideal for high data rate applications Vulnerable to interference in crowded environments	(Lonzetta et al. 2018)
WiFi (IEEE 802.11)	High data throughput suitable for multimedia and real time applications Extensive existing infrastructure Robust and standardized connectivity	High power consumption Limited effective range compared to cellular or LPWAN technologies Overhead may be excessive for simple sensor data	(Samuel 2016)
Cellular Technologies (GPRS/GSM)	Extensive global network coverage Utilizes existing infrastructure for wide coverage Reliable and mature technology	Higher power consumption compared to IoT-specific protocols Limited bandwidth and data rates Recurring subscription costs and data plan requirements	(Pereira et al. 2017)
Zigbee	Extremely low power consumption Supports mesh networking for extended range Ideal for home automation and sensor networks	Lower data rates compared to WiFi or cellular Short individual node range Can experience interference with other devices in the 2.4 GHz band	(Manpreet and Malhotra 2015)
LoRa	Very long range connectivity (up to several kilometers) Low power consumption, ideal for battery operated devices Effective in rural and remote monitoring applications	Low data rate, unsuitable for bandwidth intensive applications Limited capacity and throughput Higher latency may affect real time performance	(Centenaro et al. 2016)
LoRaWAN	Provides a complete network architecture for LoRa Scalable to support many devices over large areas Includes robust security features	Requires gateway infrastructure and network management Limited data throughput similar to LoRa More complex setup compared to point to point LoRa	(Centenaro et al. 2016)
MQTT	Lightweight messaging protocol ideal for low bandwidth IoT applications Utilizes a publish subscribe model for efficient communication Simple and scalable for many devices	Operates over an underlying network protocol (not a stand-alone transport) Security depends heavily on the broker configuration Requires a reliable broker for message distribution	(Dinculeană and Cheng 2019)
NB IoT	Excellent indoor penetration and deep coverage Very low power consumption, extending battery life Leverages existing cellular networks for wide area connectivity	Limited data rate, best suited for small data packets Higher latency compared to other IoT specific protocols	(Martinez et al. 2019)

Table 3 (continued)

Protocol	Advantages	Limitations	References
Bluetooth Low Energy (BLE)	Low power consumption ideal for wearable and sensor applications Widely supported by modern smartphones and devices Simple pairing and connection processes	Very short range compared to other protocols Limited data throughput Connection latency can be an issue for time sensitive data	(Tosi et al. 2017)
Xbee	Flexible modules that support various protocols (including Zigbee) Reliable and robust mesh networking capabilities Easy integration with embedded systems	Generally higher cost than simpler RF modules Weak Built-in Security Features Interference and signal degradation	(Haque et al. 2022)

5.2.2 Sensor calibration and drift compensation

Improving the accuracy and reliability of low-cost air quality sensors is a critical challenge, particularly as these sensors are prone to drift caused by aging, environmental changes, and cross-sensitivities. ML techniques have emerged as effective tools for addressing these issues, enabling real-time adjustments and anomaly detection to ensure precise and trustworthy data for large-scale environmental monitoring.

Table 7 illustrates the diversity of studies in this field, which span various pollutants, geographical regions, and environmental conditions. Pollutants such as CO, NO_2, O_3, and PM are commonly studied, often alongside meteorological variables like temperature and humidity.

AI-driven methods consistently outperform traditional statistical approaches, as shown in Table 8. Models such as RF, NN, and LSTM provide substantial improvements in accuracy metrics like MAE and RMSE, with some studies reporting error reductions of up to 90 percent.

Despite these successes, the field faces challenges in standardizing evaluation metrics and methodologies. Each study employs unique combinations of pollutants, reference instruments, and performance measures, complicating direct comparisons. Establishing standardized protocols would facilitate broader adoption and streamline the implementation of AI-based calibration techniques in large-scale sensor networks.

5.2.3 Anomaly detection and outlier treatment

The studies summarized in Tables 9 and 10 focus on improving anomaly detection and outlier handling in air quality monitoring. These works explore a balance between traditional statistical methods and advanced AI techniques. Statistical methods rely on predefined thresholds and distribution assumptions, making them resource-efficient for real-time monitoring in constrained environments. In contrast, ML and DL approaches excel at capturing complex temporal and spatial patterns, especially when pollutant behavior is nonlinear and nonstationary.

High-quality data is essential for training and testing in this field, but obtaining such data remains a significant challenge in real-world applications. As summarized in Table 9, each study utilized a ground truth dataset, which is crucial for evaluating the performance of anomaly detection algorithms. These datasets are typically derived from reference sensors or trusted sources, such as expert inspections or government agencies, or developed in controlled laboratories setups. They serve as benchmarks for comparing algorithm performance and identifying the most effective approach for specific applications.

Among all the applications discussed in this paper, outlier detection is the most challenging. It requires a deep understanding of the underlying data and the ability to distinguish true outliers from normal variations caused by environmental conditions.

Table 9 show that many studies focus on pollutants such as $PM_{2.5}$, PM_{10}, ozone, and nitrogen oxides, which often incorporate meteorological or indoor parameters. This diversity reflects the complexity of air quality data, which involve multiple sensor types, variable timescales, and local conditions.

Table 4 Summary of studies utilizing statistical methods, machine learning, and deep learning techniques to address data-related challenges in IoT-based air quality monitoring

Category	Techniques
Data imputation for handling missing values	
Statistical methods	ARIMA (Shaadan and Rahim 2019)
	Exponential Moving Average (Shaadan and Rahim 2019)
	Kalman Filter (Shaadan and Rahim 2019)
	Linear Interpolation (Shaadan and Rahim 2019; Junninen et al. 2004; Samal et al. 2021)
	Logistic Regression Imputation (Chen et al. 2022; Agbo et al. 2022)
	Mean Imputation (Shaadan and Rahim 2019; Samal et al. 2021; Ma et al. 2020; Kim et al. 2021; Junger and Leon 2015)
	Median Imputation (Samal et al. 2021; Chen et al. 2022; Gómez-Carracedo et al. 2014; Junger and Leon 2015)
Data imputation algorithms	Best Fit Missing Value Imputation (BFMVI) (Agbo et al. 2022)
	Expectation Maximization (EM) (Agbo et al. 2022; Junger and Leon 2015; Gómez-Carracedo et al. 2014)
	MissForest (Agbo et al. 2022)
	Multiple Imputation by Chained Equations (MICE) (Kim et al. 2021; Samal et al. 2021)
	Regression Imputation (RI) (Junninen et al. 2004)
Machine learning	Decision Trees (Ma et al. 2020)
	K-Nearest Neighbors (KNN) (Chen et al. 2022; Agbo et al. 2022; Junninen et al. 2004; Ma et al. 2020)
	Random Forests (Chen et al. 2022; Alkabbani et al. 2022)
	Self-Organizing Maps (SOM) (Junninen et al. 2004)
	Support Vector Machines (SVM) (Ma et al. 2020)
Deep learning	Autoencoders (Kim et al. 2021)
	Convolutional Denoising Autoencoder (CDAE) (Wardana et al. 2022)
	Missing Data Imputation Denoising Autoencoder (MIDIA) (Ma et al. 2020)
	Recurrent Denoising Auto-Encoder (RDAE) (Wu et al. 2022)
	Temporal Convolutional Denoising Autoencoder (TCDA) (Samal et al. 2021)
	Generative Adversarial Networks (GANs) (Wu et al. 2022)
	Bidirectional Recurrent Imputation for Time Series (BRITS) (Wu et al. 2022)
	LSTM-based Imputation (Samal et al. 2021)
Sensor calibration and drift compensation	
Statistical methods	Laboratory univariate linear regression (Zimmerman et al. 2018)
	Variational Bayesian Linear Regression (Tancev and Toro 2022)
	Markov Chain Monte Carlo (Tancev and Toro 2022)
Machine learning	Multiple Linear Regression (Zimmerman et al. 2018; Ali et al. 2021; Park et al. 2021)
	Random Forest Regression (Zimmerman et al. 2018; Bush et al. 2022)
	Random Forest machine learning model (Zimmerman et al. 2018)
	Empirical multiple linear regression (Zimmerman et al. 2018)
Deep learning	Artificial Neural Network (Ali et al. 2021)
	LSTM (Park et al. 2021)
	Deep Neural Network (Park et al. 2021)
	Variational Bayesian Neural Network (Tancev and Toro 2022)
Anomaly detection and outlier treatment	

Table 4 (continued)

Category	Techniques
Statistical methods	Multivariate Linear Regression (Ottosen and Kumar 2019)
	Control Charts & Thresholding
	Z-score (Mahajan et al. 2020)
	InterQuartile Range (IQR) (Mahajan et al. 2020)
	Box–Cox Transformation (Martínez et al. 2014)
	Bootstrap Thresholding (Martínez et al. 2014)
	Outlier Detection Tests
	Grubb's Test (Mahajan et al. 2020)
	Tietjen-Moore Test (Mahajan et al. 2020)
	Hampel's Test (Mahajan et al. 2020)
	Change Point Detection
	Pruned Exact Linear Time (Ottosen and Kumar 2019)
	Time Series Modeling
	ARIMA (Wang et al. 2020)
	HI-ARIMA (Wang et al. 2020)
	SARIMA (Maltare and Vahora 2023)
Machine learning	Support Vector Machine (SVM) (Maltare and Vahora 2023; Park et al. 2023; Jesus et al. 2021)
	k-Nearest Neighbors (k-NN) (Ottosen and Kumar 2019)
	Extreme Learning Machine (ELM) (Wang et al. 2020)
	Nearest Cluster Predictor (Hill and Minsker 2010)
	Multilayer Perceptron (MLP) (Hill and Minsker 2010)
	Artificial Neural Network (Ottosen and Kumar 2019; Hill and Minsker 2010; Jesus et al. 2021)
	Majority Voting Ensemble (Rollo et al. 2023)
	Latent Feature Clustering (DBSCAN) (Park et al. 2023)
Deep learning	LSTM Autoencoder (Wei et al. 2023)
Air quality index (AQI) estimation	
Statistical methods	Linear Regression (LR) (Hossain et al. 2021)
	Seasonal Auto-Regressive Integrated Moving Average (SARIMA) (Maltare and Vahora 2023)
Machine learning	Decision Tree (DT) (Hossain et al. 2021)
	K-Nearest Neighbors (KNN) (Hossain et al. 2021)
	Random Forest Regression (RFR) (Liu et al. 2019; Maltare and Vahora 2023; Hossain et al. 2021)
	Support Vector Machines (SVM) (Maltare and Vahora 2023; Liu et al. 2019; Hossain et al. 2021)
	Support Vector Regression (SVR) (Janarthanan et al. 2021; Liu et al. 2019; Kok et al. 2017; Maltare and Vahora 2023)
Deep learning	Artificial Neural Networks (ANN) (Hossain et al. 2021)
	Gated Recurrent Unit (GRU) (Hossain et al. 2021)
	Hybrid CNN-LSTM (Hossain et al. 2021)
	Hybrid GRU-LSTM (Hossain et al. 2021)
	Long Short-Term Memory (LSTM) (Kok et al. 2017; Janarthanan et al. 2021; Hossain et al. 2021)
	LSTM + SVM + SARIMA (Maltare and Vahora 2023)
	LSTM + SVR Hybrid (Janarthanan et al. 2021)
	LSTM with Deep Neural Networks (LSTM-DNN) (Kim et al. 2022)
Short-term air quality forecasting	

Table 4 (continued)

Category	Techniques
Statistical methods	ARIMA (Freeman et al. 2018; Li et al. 2016; Zhang and Zhang 2023; Xu and Yoneda 2021)
	Linear Regression (Navares and Aznarte 2020)
	Multiple Linear Regression (Freeman et al. 2018)
	Principal Component Analysis (Freeman et al. 2018)
Machine learning	Decision Trees (Freeman et al. 2018)
	Gradient Boosted Regression Trees (Li et al. 2022)
	K-means (Wang et al. 2022)
	Random Forest (Navares and Aznarte 2020; Zhang and Zhang 2023)
	Support Vector Machine (Wang et al. 2022; Freeman et al. 2018)
	Support Vector Regression (Du et al. 2021; Li et al. 2016; Zhang and Zhang 2023)
	XGBoost (Zhang and Zhang 2023; Cui et al. 2023)
Deep learning	Convolutional Neural Networks (CNN)
	1D-CNN (Zhang and Zhang 2023)
	2D-CNN (Wang et al. 2022)
	CNN-LSTM (Chen et al. 2022)
	Graph Neural Networks (Liang et al. 2023)
	Long Short-Term Memory (Du et al. 2021; Freeman et al. 2018; Li et al. 2016; Navares and Aznarte 2020; Zhang and Zhang 2023; Cui et al. 2023; Xu and Yoneda 2021; Chen et al. 2022)
	LSTM Autoencoder (Xu and Yoneda 2021)
	Transformer Networks (Liang et al. 2023; Zhang and Zhang 2023; Cui et al. 2023; Chen et al. 2022)
	CNN-Transformer Hybrid Model (Chen et al. 2022)
	Double Output Vision Transformer (Wang et al. 2022)
	Sparse Attention-based Transformer (Zhang and Zhang 2023)
	TSF-Transformer (Li et al. 2022)
	Vision Transformer (Wang et al. 2022)

As shown in Table 10, combining anomaly detection with gap-filling techniques, such as interpolation, spline fitting, or model-based reconstruction, typically results in more complete and reliable datasets.

Deep learning models, including autoencoders and LSTM networks, stand out in Table 10 for their ability to learn complex temporal dependencies and nonlinear relationships. Autoencoders, for example, use reconstruction loss as an effective anomaly signal, especially when combined with other algorithms. LSTM autoencoders integrate memory cells to detect unusual events in sequential data, achieving high accuracy and low false positives in long-term indoor CO_2 datasets.

Despite their advantages, data-driven models require significant computational resources and high-quality training datasets for optimal performance. Furthermore, their effectiveness can decrease if they are not regularly updated with recent sensor data, especially in dynamic environments where pollution patterns evolve. A practical solution is to implement routine retraining and calibration cycles, especially for low-cost sensor networks prone to drift.

Table 5 Summary of parameters used in the studies for data imputation

Study	Target variable	Support variables	Location	Timespan	Metrics	Sampling rate	Missing rates (%)
(Wu et al. 2022)	CO_2, RH, PM_{10}, $PM_{2.5}$, T, VOC		Gainesville, USA	2020	MRE, MAE	10-minute	5, 25, 50, 75
(Wardana et al. 2022)	NO_2, PM_{10}, $PM_{2.5}$, CO, O_3		London (UK), Delhi (India), Beijing (China)	2013-2021	RMSE, MAE, R^2, RIR	Hourly	20, 40, 60, 80
(Shaadan and Rahim 2019)	PM_{10}		Shah Alam, Malaysia	2015	RMSE, MAE, R^2, AI	Hourly	5, 10, 15
(Samal et al. 2021)	$PM_{2.5}$	Temperature, pressure, rainfall, wind direction, wind speed	Beijing, China	2013-2017	RMSE, MAE, MAPE	Hourly	20, 30, 40, 50
(Ma et al. 2020)	CO, NMHC, NO_x, NO_2, O_3, T, RH, AH		Italy	2004-2005	RMSE, Macro-F	Hourly	5, 10, 50
(Kim et al. 2021)	$PM_{2.5}$, PM_{10}		Guro-gu, Seoul, and Dangjin-si, South Korea	2020-2021	MAE, sMAPE	Minute	7.91, 16.1, 20
(Junninen et al. 2004)	NO_x, NO_2, O_3, PM_{10}, SO_2, CO	Wind speed, wind direction, temperature, relative humidity	Belfast, UK and Helsinki, Finland	1998	d_2, R^2, RMSE, MAE	Hourly	10, 25
(Junger and Leon 2015)	PM_{10}	Temperature, relative humidity	São Paulo, Brazil	2004	RMSE, MAE, BIAS, PV, Pearson's r, d_2	Daily	5, 10, 20, 30, 40
(Gómez-Carracedo et al. 2014)	NO, NO_2, NO_x, CO, O_3, PM_{10}, $PM_{2.5}$, PM_1		A Coruña, Spain	2006, 2009, 2010	Factor analysis, Variance explained by factors, Comparison of imputed values' distributions	Hourly	23.5, 11.95, 3.85
(Chen et al. 2022)	$PM_{2.5}$	SO_2, NO_2, O_3, CO, PM_{10}	Lanzhou, China	2019	MAE, RMSE, MAPE	Hourly	5, 10, 20, 40
(Alkabbani et al. 2022)	AQI, $PM_{2.5}$, PM_{10}, O_3, SO_2, NO_2, CO	Temperature, wind speed, wind direction, relative humidity	Al-Jahra City, Kuwait	2013–2015	MAE, RMSE, MSE, R^2	Hourly	10.96, 10.36, 8.01
(Agbo et al. 2022)	C_6H_6	Temperature, relative humidity, CO, NMHC, NO_x, O_3	Italy	2004-2005	RMSE, MAE, R^2	Hourly	10, 20, 30, 40

The table summarizes the target and support variables, location, timespan, metrics, sampling rate, and missing rates for various studies focused on data imputation in air quality monitoring

5.2.4 Air quality index estimation

AQI estimation is a key component of air quality monitoring, providing a standardized measure that translates pollutant concentrations into a user-friendly scale for public awareness. This estimation process typically involves applying statistical or AI-based techniques to predict the AQI value from measured concentrations of pollutants and relevant meteorological features.

A reliable ground-truth dataset is crucial for AQI estimation, just as it is for outlier detection tasks. As summarized in Table 11, most studies employ trusted AQI data obtained from government agencies or professional monitoring equipment. These high-quality datasets ensure that models are trained on accurate targets, reducing bias and improving reliability. In this context, effective data preprocessing steps (such as outlier removal, missing data imputation, and feature scaling) also help to enhance model performance.

Various regression models, including traditional machine learning methods and modern deep learning techniques, have been explored for AQI prediction. Their performance is typically evaluated using standard metrics such as RMSE, MAE, and R^2. Table 12 shows that DL methods, especially LSTM networks, consistently achieve superior results. This superiority largely arises from their ability to capture temporal dependencies in sequential AQI data, making them well suited for forecasting tasks in dynamic environments.

Additionally, studies including meteorological features alongside pollutant concentrations report better accuracy than those relying on pollutant data alone, emphasizing the influence of weather conditions on air quality. Beyond single model approaches, hybrid methods that combine LSTM with other ML or DL models have outperformed individual techniques in many cases.

As was the case with outlier detection, the AQI estimation task also faces challenges related to the availability of high-quality datasets. Many studies rely on government-provided data, which may not be accessible in all regions.

This limitation can hinder the development and deployment of effective AQI estimation models in areas with limited resources or infrastructure.

5.2.5 Short-term air quality forecasting

In short-term air quality forecasting, the objective is to predict future pollutant concentrations based on historical data, providing projections for the next hours or days. This domain has seen significant advancements with the adoption of ML and DL techniques, particularly architectures like LSTM, which are well-suited for time series data, and attention mechanism-based models, such as transformers, which excel at handling sequential data with long-range dependencies. Nevertheless, statistical methods like ARIMA remain widely used, as shown in Table 4.

Table 13 summarizes the reviewed studies. Unlike previous sections, all studies incorporate both pollutant and meteorological features, a common practice in this field. Data sources are consistent, with most studies relying on government-provided datasets. The sampling rate is generally more sparse, with data collected daily or hourly, compared to the minute or second intervals observed in other sections. Evaluation metrics remain similar, with RMSE and MAE being the most frequently used.

Table 6 Summary of reviewed studies on missing data imputation, including imputed variables, missing data rates, applied techniques, and a comparison of evaluation metrics across studies

Study	Tested technique	Specific technique	Imputed variable	Evaluation metrics							
				RMSE				MAE			
				Tested missing rates (%)							
				5	10	20	25	5	10	20	25
(Wardana et al. 2022)	Deep Learning	Convolutional Denoising Autoencoder	NO_2			7.32				4.07	
			CO			471.88				296.36	
			O_3			8.91				3.93	
			$PM_{2.5}$			15.49				9.14	
			PM_{10}			5.24				3.87	
(Shaadan and Rahim 2019)	Statistical Method	Kalman Filter & ARIMA	PM_{10}	9.43	9.55			7.08	7.16		
(Samal et al. 2021)	Deep Learning	Autoencoder (TCDA)	$PM_{2.5}$			20.00				19.00	
(Ma et al. 2020)	Deep Learning	Autoencoder (MIDIA)	Not specified*	0.03	0.03	0.03	0.03				
(Junninen et al. 2004)	Machine Learning	avg. of kNN, SOM and MLP	NO_x		24.20		22.70		17.30		15.40
(Junger and Leon 2015)	Data imputation algorithm	Expectation Maximization (EMeMV)	PM_{10}	0.23				0.16			
(Agbo et al. 2022)	Data imputation algorithm	Best Fit Missing Value Imputation (BFMVI)	C_6H_6		0.0118	0.0290			0.0006	0.0019	
(Chen et al. 2022)	Statistical Method	Logistic Regression Imputation (FTLRI)	$PM_{2.5}$	6.47	7.51	6.48		4.32	5.68	5.02	
(Wu et al. 2022)	Deep Learning	Generative Adversarial Neural Network (IM-GAN)	Not specified+					9.55			13.90
								2.78			2.55
				6.70	**7.89**	**8.01**	**~10**	**6.70**	**7.89**	**8.01**	**~10**

Table 6 (continued)

Study	Tested technique	Specific technique	Imputed variable	Evaluation metrics							
				RMSE				MAE			
				Tested missing rates (%)							
				5	10	20	25	5	10	20	25
(Alkabbani et al. 2022)	Machine Learning	Random Forest-based imputation (MissForest) and Artificial Neural Network	O_3				5.58				4.55
			SO_2		13.82				6.01		
			CO	0.2110				0.1690			
			NO_2				12.67				9.99
			$PM_{2.5}$				3.76				2.78
			PM_{10}			21.96				7.98	
				26.03		**28.96**		**26.03**		**28.96**	
(Kim et al. 2021)	Deep Learning	Autoencoder	$PM_{2.5}$	2.74							
			PM_{10}			1.15					

* This study presented a table with results, but the imputed variable was not specified. + This study presented results on two different datasets, GAMS and Gainesville, but no specific imputed variable was mentioned with the results

Table 7 Summary of parameters used in the reviewed studies air quality sensor calibration studies, including target pollutants, meteorological variables, data sources, geographical contexts, time periods, sampling frequencies, and performance evaluation metrics

Study	Pollutants and meteorological features	Data source	Country	Timespan	Sampling rate	Evaluation metrics
(Zimmerman et al. 2018)	CO, NO_2, CO_2, O_3; T, RH	RAMP sensor deployments at Carnegie Mellon University and Allegheny County Health Department	USA	2016-2017	15-minute averages	MAE, Pearson r, CvMAE, RMSE, MBE, R^2, CRMSE
(Bush et al. 2022)	NO2, PM_{10}, $PM_{2.5}$; T, RH	Defra Oxford St Ebbe's AURN monitoring station	UK	2020	15 min	MAE, R^2
(Ali et al. 2021)	CO, NO_2, PM_{10}; T, RH	Auckland City Council AQM station	New Zealand	2019	1 min (raw), 10 min (averaged)	MAPE, R^2, RMSE
(Park et al. 2021)	$PM_{2.5}$; Temperature, Humidity	Sensirion SPS30 and SHT85 sensors, TEOM gravimetric instrument	South Korea	2019-2020	Hourly (time-series)	R^2, RMSE
(Tancev and Toro 2022)	CO, NO_x, NO_2, C_6H_6; T, RH	Publicly available dataset	Italy	1 year	1 sample per hour	R^2, Negative log likelihood

Table 8 Summary of reviewed studies sensor calibrarion, including imputed variables, missing data rates, applied techniques and a comparison of evaluation metrics across studies

Study	Tested technique	Specific technique	Notes
(Park et al. 2021)	Deep Learning	Hybrid LSTM	Proposed as a state-of-the-art model for PM2.5 sensor calibration. Metrics: RMSE improved by 41–60% over raw data, 30–51% over MLR, and 8–40% over DNN; $R^2 = 0.93$
(Ali et al. 2021)	Deep Learning	Artificial Neural Network	Demonstrates the use of ANN for low-cost sensors with LoRaWAN connectivity. Metrics: MAPE = 38.89%, $R^2 = 0.78$
(Bush et al. 2022)	Machine Learning	Random Forest Regression	Field-deployed for a 7-month period alongside reference instrumentation, showing significant improvement in accuracy. Metrics: Improved MAE by 37–94%; expanded uncertainty = 29% for NO_2, 21% for PM_{10}, and 27% for $PM_{2.5}$
(Zimmerman et al. 2018)	Machine Learning	Random Forest	Robust performance validated over 16-week testing periods. Metrics: MAE = CO 38 ppb (14%), CO_2 10 ppm (2%), NO_2 3.5 ppb (29%), O_3 3.4 ppb (15%); Pearson's r > 0.8
(Tancev and Toro 2022)	Deep Learning	Variational Bayesian Neural Network	Presents Bayesian approaches for improved for predictive maintenance with enhanced uncertainty estimation and anomaly detection capability. Metrics: Numerical metrics not explicitly provided, only charts presented

Table 9 Comparison of representative studies on anomaly detection and outlier treatment in air quality monitoring, with emphasis on differences in pollutants measured, sampling strategies, data sources and validation approaches

Study	Pollutants and meteorological features	Data source	Country	Timespan	Sampling rate	Evaluation metrics	Ground truth
(Hill and Minsker 2010)	Windspeed	Shoreline Environmental Research Facility (SERF)	USA	2004	1 sample per second	False Positive Rate, False Negative Rate	Manual inspection by SERF experts
(Jesus et al. 2021)	None	SATURN (CMOP Science and Technology University monitoring network)	USA	Training: 2009-2010; Testing: 2013-2014	1 measurement every 6 min	Detection Rate, False Positive Rate, ROC curve analysis	Expert inspection of testing dataset
(Mahajan et al. 2020)	PM_{10}, $PM_{2.5}$	Indian air quality sensor networks	India	2010-2015	Not explicitly stated	Detection accuracy, Execution time, Incremental vs. full dataset performance	Known outliers in benchmark datasets; real-world data assumed clean
(Wei et al. 2023)	CO_2	SKOMOBO (School Monitoring Box) deployment by Massey University and NIWA	New Zealand	2018	1-minute interval	Accuracy, Precision, Recall, F1-score, AUC-ROC	Labeled based on 2-sigma rule (threshold of CO2 values)
(Wang et al. 2020)	None	Local municipal air quality monitoring stations	China	2017-2018	Daily	MAE, RMSE, MAPE, Pearson Correlation Coefficient, Index of Agreement	Observed AQI values from official monitoring stations
(Rollo et al. 2023)	NO, NO_2, O_3, CO	TRAFAIR low-cost air quality sensor network	Italy	2019-2020	10-minute aggregation	RMSE, Accuracy, Precision, Recall, F1-Score, MAPE	Sensor status logs and synthetic anomaly injection
(Park et al. 2023)	$PM_{2.5}$, PM_{10}, TVOC, CO_2, CO, CH_2O	Custom environmental sensor network (IoT-based)	Korea	Two weeks of data for lab testing	Every 2 min	Accuracy, Reconstruction Error, SVM Score	Laboratory testing with manually injected abnormal events
(Ottosen and Kumar 2019)	NO, NO_2, SO_2, CO, O_3	AQMesh low-cost sensor (Environmental Instruments Ltd)	UK	2017-2018	Every 15 min	RMSE, MAE, Correlation, Index of Agreement	Unsupervised thresholding using statistical change detection and residual analysis
(Martínez et al. 2014)	CO, NO_2, SO_2	Government of Asturias– Air Quality Monitoring Stations	Spain	2006-2011	Every 15 min	Number of outliers detected, Detection rates comparison	Comparative validation between classical and functional detection methods

Table 9 (continued)

Study	Pollutants and meteorological features	Data source	Country	Timespan	Sampling rate	Evaluation metrics	Ground truth
(Maltare and Vahora 2023)	$PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3, NH_3, Pb, Ni, As, Benzo(a)pyrene, Benzene	Central Pollution Control Board (CPCB) and SAFAR	India	2015-2021	Hourly	R^2 Score, MSE, RMSE, MAE	Actual AQI values from CPCB and SAFAR

Table 10 Summary of the best-performing models and key findings from the reviewed studies on anomaly detection and outlier treatment

Study	Tested technique	Specific technique	Notes
(Hill and Minsker 2010)	Machine Learning	Multilayer Perceptron	Best-performing model achieved a 1% false positive rate and 2% false negative rate. The experimental setup included 10-fold cross-validation. Anomalies were primarily caused by sensor or transmission faults
(Jesus et al. 2021)	Machine Learning	Artificial Neural Network	Achieved 100% detection rate for most sensors at a 0.998 threshold. False positive rate as low as 0% depending on threshold selection. Validated on real-world sensor data from the Columbia River estuary
(Mahajan et al. 2020)	Statistical Methods	Hampel's Test and IQR	High accuracy with low execution time, suitable for real-time use. Incremental methods showed negligible performance loss compared to full dataset analysis
(Maltare and Vahora 2023)	Machine Learning	Support Vector Machine	Achieved $R^2 = 0.9989$ and RMSE = 4.94 for AQI prediction. Preprocessing included outlier removal and feature selection
(Martínez et al. 2014)	Statistical Method	H-Modal Depth	Identified fewer but more relevant outliers compared to classical methods. Robust to measurement error and seasonal effects
(Ottosen and Kumar 2019)	Statistical Methods	Linear Interpolation	Linear interpolation outperformed other methods for short gaps. Combined outlier detection with gap filling for low-cost sensors
(Park et al. 2023)	Machine Learning	LSTM Autoencoder	Achieved 97.66% accuracy in detecting anomalies. Combined sensitivity of LSTM-AE with precision of OC-SVM
(Rollo et al. 2023)	Machine Learning	AIrSense with anomaly	Reduced NO_2 RMSE from 53.13 to 8.67. Combined anomaly detection with VAR-based repairing for calibration. Deployed in Modena, Italy
(Wei et al. 2023)	Machine Learning	LSTM-Autoencoder	Achieved 99.50% accuracy and 94.68% F1-score for CO_2 anomaly detection. Designed for low-cost IAQ monitoring in schools
(Wang et al. 2020)	Machine Learning	Extreme Learning Machine	Achieved lowest error rates and highest correlation for AQI prediction. Combined outlier detection, signal decomposition, and machine learning

Table 14 summarizes the results of the reviewed studies. DL models consistently outperform traditional ML and statistical methods, with transformers and LSTM-based models achieving the best results across all studies.

Both LSTM and transformers effectively capture temporal dependencies in the data and have been extensively studied in various research fields, including air quality forecasting. The use of attention mechanisms in transformers allows them to focus on relevant parts of the input data, enhancing their performance in long-range forecasting tasks.

Among all the fields discussed in this paper, short-term air quality forecasting benefits from easier data availability, as no ground truth is required. This is because the goal is to predict future values based on historical data, with models trained on past observations. However, this also makes the models more sensitive to noise and outliers in the data, which can negatively impact their performance. Therefore, it is crucial to carefully preprocess the data and remove any anomalies before training the models.

Table 11 Representative studies on AQI estimation, comparing pollutant inputs, data sources, sampling rates, and evaluation strategies across different geographic and methodological contexts

Study	Pollutants and meteorological features	Data source	Timespan	Sampling rate	Evaluation metrics	Ground truth AQI
(Hossain et al. 2021)	$PM_{2.5}$, PM_{10}, NO_2, CO, SO_2, O_3, T, RH, Wind speed	Central Pollution Control Board (CPCB), Delhi, India	2018-2021	Hourly	MAE, RMSE, R^2	Official AQI values published by CPCB monitoring stations in Delhi
(Maltare and Vahora 2023)	$PM_{2.5}$, PM_{10}, CO, O_3, NO_2, SO_2, NH_3, Pb, Ni, As, Benzo(a) pyrene, Benzene	Central Pollution Control Board (CPCB) and SAFAR, Ahmedabad, India	2015-2021	Hourly	R^2, MSE, RMSE, MAE	AQI derived using Indian CPCB formula based on pollutant sub-indices
(Liu et al. 2019)	$PM_{2.5}$, O_3, SO_2, PM_{10}, NO_2, CO, NO_X, non-methane hydrocarbons, benzene	Beijing Municipal Environmental Monitoring Center and Italian dataset	2013-2018 (Beijing); 2004-2005 (Italy)	Hourly	RMSE, R^2, Correlation Coefficient (r)	Hourly averaged AQI from certified analyzers
(Kok et al. 2017)	O_3, NO_2, PM, CO, SO_2	CityPulse EU FP7 Project, Aarhus (Denmark) and Brasov (Romania)	2013-2015	5-minute	RMSE, MAE, Precision, Recall, F1-Score, Accuracy	Environmental Protection Agency (EPA) AQI standard with three threshold levels
(Janarthanan et al. 2021)	$PM_{2.5}$, NO_2, SO_2, CO, O_3, RH, atmospheric pressure, wind speed, wind degree	Central Pollution Control Board (CPCB), Chennai, India	2019-2020	15-minute	RMSE, R^2, MSE	Official AQI values published by CPCB monitoring stations in Chennai

Table 12 Summary of AQI estimation models with their learning approaches, predictive techniques, and performance metrics, presenting variation in methodological complexity and accuracy across studies

Study	Tested technique	Specific technique	Notes	RMSE	MAE	R2	MSE
(Hossain et al. 2021)	Deep Learning	Hybrid GRU-LSTM	AQI prediction using time series modeling with pollutant and meteorological features	0.07	0.05		0.01
(Maltare and Vahora 2023)	Machine Learning	Support Vector Machine (RBF Kernel)	AQI prediction using regression on pollutant concentration features	4.94		1.00	24.44
(Liu et al. 2019)	Machine Learning	Support Vector Regression (SVR)	AQI and NOx concentration prediction using regression with environmental sensor data	7.67		0.98	
(Kok et al. 2017)	Deep Learning	Long Short-Term Memory (LSTM)	AQI prediction using regression with pollutant concentration features	3.26	2.81		
(Janarthanan et al. 2021)	Hybrid Model	SVR with LSTM	AQI prediction using regression with pollutant and meteorological features			0.82	0.18

Table 13 Comparison of short-term air quality forecasting studies based on data sources, input features, temporal resolution, and evaluation metrics

Study	Pollutants and meteorological features	Data source	Country	Timespan	Sampling rate	Evaluation metrics
(Freeman et al. 2018)	O_3, NO_x, SO_2, CO, VOCs, PM, wind speed, wind direction, T, RH, solar radiation, atmospheric pressure	OPSIS differential optical absorption spectroscopy analyzers, Kuwait	Kuwait	2012–2014	Hourly	MAE, RMSE
(Navares and Aznarte 2020)	CO, NO_2, O_3, PM_{10}, SO_2, pollen, T, RH, wind speed, rainfall, atmosferic pressure	Madrid Municipal Air Quality Monitoring Grid & Spanish Aerobiological Network	Spain	2001–2013	Daily	RMSE
(Chen et al. 2022)	O_3, NO, NO_2, SO_2, CO, wind speed, wind direction, T	14 monitoring stations from Beijing Municipal Ecological and Environmental Monitoring Center	China	2014–2021	Daily	RMSE, NRMSE, MAE
(Zhang and Zhang 2023)	$PM_{2.5}$, dew point, T, atmosferic pressure, wind direction, wind speed, rain/snow, RH	Beijing and Taizhou $PM_{2.5}$ datasets	China	2010–2014 (Beijing), 2017–2019 (Taizhou)	Hourly	RMSE, MAE, R^2
(Xu and Yoneda 2021)	$PM_{2.5}$, T, atmosferic pressure, RH, wind direction, wind speed, weather conditions	18 city-wide $PM_{2.5}$ monitoring stations, Beijing	China	2017–2018	Hourly	RMSE, MAE, SMAPE
(Liang et al. 2023)	$PM_{2.5}$, PM_{10}, NO_2, CO, O_3, SO_2, weather conditions, T, RH, wind speed, wind direction	Nationwide air quality and meteorological datasets, Chinese mainland	China	2015–2018	Every 3 h	MAE, RMSE
(Cui et al. 2023)	$PM_{2.5}$, SO_2, NO_2, PM_{10}, CO, O_3, T, atmosferic pressure, dew point, rainfall, wind speed, wind direction	China Meteorological Data Network, Beijing	China	2013–2017	Hourly	EVS, R^2, MAE, MSE, MAPE

5.3 Hardware constraints

The adoption of low-cost sensors has become increasingly popular in recent years, offering a more affordable alternative to traditional air quality monitoring methods. These sensors are typically compact and lightweight, making them well-suited for deployment in diverse environments such as urban areas, industrial sites, and remote locations. Often, they are interfaced with microcontrollers or single-board computers, which provide the processing power and connectivity required for efficient data collection and transmission.

However, the use of low-cost sensors also introduces several challenges. They frequently exhibit limitations in accuracy, precision, and long-term stability, and their performance can be adversely affected by environmental factors such as temperature, humidity, and pressure. Additionally, the lack of standardized evaluation procedures complicates the comparison of sensor performance across different studies.

Section 4 of this paper focuses on the application of AI to air quality monitoring, including techniques for sensor calibration, data imputation, and anomaly detection. Many of the issues addressed by these AI techniques originate from inherent hardware shortcomings, which may cause sensors to misread data, fail to transmit information properly, or produce inaccurate environmental measurements. These AI methods are specifically designed to mitigate such issues, acknowledging that hardware failures and data corruption are, to some extent, inevitable.

However, most techniques discussed in previous sections overlook the hardware used for data measurement and assume cloud-based processing. These techniques are generally designed for execution on servers or cloud systems, where processing power and memory are ample.

By contrast, Table 1 summarize hardware platforms and communication protocols commonly used in air quality monitoring. Most of these microcontrollers and single-board computers are cost-effective and low-power, making them suitable for remote deployments. Table 2 shows that many reported systems are microcontroller-based, which limits CPU, RAM, and storage resources. Such systems often run only lightweight RTOS or have no OS at all, and high-level programming languages like Python are typically unsupported.

While microcontroller-based devices can support AI techniques, their constrained resources limit both data processing capabilities and the complexity of models they can handle. Frameworks such as TensorFlow Lite (David et al. 2021) and TinyML (Dutta and Bharali 2021) enable lightweight AI implementations on these devices, but the trade-off is reduced model sophistication. On the other hand, platforms built around microprocessors, like the Raspberry Pi or AR9331, offer much higher computational capabilities, support advanced programming languages, and are equipped to operate complete operating systems such as Linux. This enables them to execute machine learning algorithms in a manner more akin to a computer. These advantages make them better suited for more complex machine learning tasks, albeit at the cost of higher energy consumption and increased expense.

5.4 Edge vs. cloud processing

Given the hardware constraints described earlier, it is important to balance the use of edge-based and cloud/server-based processing in air quality monitoring systems. Balancing edge and cloud processing is essential in air quality monitoring systems due to hardware limita-

Table 14 Summary of AQI estimation models with their learning approaches, predictive techniques, and performance metrics, showing variation in methodological complexity and obtained results across studies

Study	Tested technique	Specific technique	Notes	Evaluation metrics	
				RMSE	MAE
(Chen et al. 2022)	Deep Learning	Hybrid CNN-Transformer	Combines CNN for local feature extraction and Transformer for global sequence learning	7.75	5.92
(Cui et al. 2023)	Deep Learning	Transformer	Uses attention mechanisms for hourly $PM_{2.5}$ prediction. Incorporates STL decomposition and sliding window segmentation	3.28	-
(Liang et al. 2023)	Deep Learning	AirFormer	Employs Dartboard Spatial MSA (DS-MSA) for spatial dependencies and Causal Temporal MSA (CT-MSA) for temporal dependencies. Predicts $PM_{2.5}$ for 1-72 h ahead	32.36	16.03
(Xu and Yoneda 2021)	Deep Learning	LSTM-Autoencoder	Encodes meteorological data with stacked autoencoders and predicts $PM_{2.5}$ concentrations		
(Zhang and Zhang 2023)	Deep Learning	Sparse Attention-based Transformer (STN)	Excels in short- and long-term $PM_{2.5}$ predictions. Reduces time complexity from $O(L^2)$ to $O(L \ln L)$	19.04	11.13
(Navares and Aznarte 2020)	Deep Learning	GP-LSTM	Groups pollutants by class for improved accuracy. Best for CO, NO_2, and PM_{10} predictions	0.083	
(Freeman et al. 2018)	Deep Learning	RNN-LSTM	First application of LSTM for air pollution prediction. Handles missing data with novel imputation techniques	0.8	0.41

tions. Edge processing analyzes data directly on sensors or local gateways, which reduces latency, cuts data-transfer costs, and keeps sensitive information on site. Its main drawback is limited processing power, which restricts the complexity of models that can be deployed.

Cloud processing uses scalable resources to run advanced ML and DL models. However, this approach comes with higher data-transfer costs, potential delays, and increased security risks.

Tables 1 and 3 list common hardware platforms and communication protocols. Devices that support high-level programming languages and have sufficient processing power can run models locally, making edge processing a viable option. In contrast, microcontroller-based platforms with limited resources often rely on cloud solutions.

Communication protocols also affect data transfer. Low-power protocols such as LoRaWAN, Zigbee, and NB-IoT provide long-range connectivity but have limited bandwidth and higher latency, making them more suitable for remote deployments rather than real-time processing. GSM and LTE offer better bandwidth and lower latency at a higher cost, while GPRS is appropriate for infrequent transmissions. In time-sensitive scenarios, such as emergencies or areas with poor connectivity, local processing can reduce delays and costs.

Most air quality monitoring systems favor cloud processing because it supports complex models and integrates easily with various data sources for long-term monitoring. Edge processing should be used only when immediate data analysis is needed, to ensure that devices

have enough processing power, reliable energy, and protection against environmental challenges like humidity and overheating.

5.5 Policy impacts of IoT air quality systems

In recent years, growing concern about air pollution and its impact on public health has driven increased interest in IoT-based air quality monitoring systems. These systems utilize low-cost sensors, advanced data processing techniques, and real-time communication protocols to provide accurate and timely air quality information.

Numerous examples worldwide demonstrate how these systems have contributed to actionable environmental policies. Mahajan et al. provided a comprehensive review of citizen-generated air quality data and its role in shaping policy in various regions (Mahajan et al. 2022). The review highlighted several impactful air quality monitoring projects, organized below by continent:

- **Asia:**
  - Air Box (Chen et al. 2017)
  - Safecast (Brown et al. 2016)
- **Africa:**
  - Open Seneca (open-seneca)
  - OpenAQ (Hasenkopf et al. 2015)
  - GH Air (Sewor et al. 2021)
- **Europe:**
  - Citisense (Nikzad et al. 2012)
  - Air Kit (Mahajan et al. 2021)
  - hackAIR (Kosmidis et al. 2018)
- **Latin America:**
  - Ciudadanos Científicos (Hoyos et al. 2019)
  - CanAirIO (Bernal et al. 2020)
  - Open Seneca

These projects vary in scope, from crowdsourced air quality data collection to large-scale deployments such as Safecast, which established a sensor network to map radiation levels around the Fukushima Daiichi Nuclear Power Plant.

Although many IoT-based citizen science initiatives have focused primarily on public awareness and data generation, only a few have demonstrated direct policy influence. As noted by Mahajan et al., tangible policy impacts have been observed in select projects in Taiwan, Japan, Kenya, and Belgium. These examples highlight the potential of citizen-

generated data from IoT systems to inform and shape environmental policy, especially in regions where traditional government monitoring infrastructure is limited or absent.

6 Conclusion

The integration of AI techniques into IoT-based air quality monitoring systems has enhanced their performance by enabling real-time analysis and predictive modeling capabilities.This paper reviews recent advances in air quality monitoring, emphasizing the importance of state-of-the-art AI models such as convolutional neural networks (CNNs), LSTM networks, transformer architectures, and autoencoders. These models address common issues such as data reliability, sensor calibration, missing data management, and anomaly detection, thereby improving the accuracy and effectiveness of monitoring systems.

A common trend in recent research is using low-cost, commercially available sensors, which offer an affordable alternative for large-scale deployment. However, these sensors still exhibit shortcomings, particularly in their sampling capacity, reliability, and long-term stability. There is no standardized evaluation procedure to properly assess their performance under laboratory conditions, leading to inconsistencies, as each study employs its own methodology. Establishing a unified framework for short-term, medium-term, and long-term performance evaluation would ensure consistency and reliability in air quality measurements.

To move forward, the future of air quality monitoring depends on developing intelligent, modular systems that integrate hybrid AI models, edge computing, and multimodal data sources. Addressing power constraints, communication infrastructure, and privacy concerns will be critical to deploying these systems in remote and heterogeneous environments. Continued interdisciplinary collaboration between academia, industry, and policymakers will be vital to overcoming current limitations and achieving widespread adoption of AI and IoT-enabled environmental monitoring and public health protection solutions.

The findings of this review have important implications for researchers, engineers, and environmental agencies. By providing a structured overview of AI techniques and their application areas, along with current challenges and technology trends, this study offers a foundation for developing more intelligent, scalable, and context-aware air quality monitoring systems. Furthermore, the classification framework and identifying gaps can help prioritize future research efforts, encourage standardization in system design, and support evidence-based policymaking to improve air quality and public health outcomes.

7 Limitations

Despite significant progress, AI and IoT-based air quality monitoring systems face several challenges that limit their scalability and reliability. Data bias and region-specific pollutant variations reduce the model's generalizability, while sensor drift, cross-sensitivity, and environmental changes affect long-term accuracy. These issues call for advanced calibration techniques and adaptive learning algorithms. However, sensor calibration can be costly and labor-intensive, especially at scale.

Privacy concerns also arise when integrating geolocated or multimodal data, requiring robust data protection strategies. The computational demands and energy consumption of

DL models limit their use in resource-constrained settings. While edge computing offers real-time processing and improved reliability, it depends on a stable power supply. Cloud-based solutions remain more viable in remote areas, though with higher latency.

Scaling to national or multinational levels presents challenges related to data standardization, device interoperability, and infrastructure coordination. Addressing these limitations requires continued interdisciplinary collaboration to build scalable, adaptable, and privacy-respecting solutions for real-world deployment.

8 Future directions

- Future research should explore hybrid modeling approaches that combine multiple ML techniques to improve system robustness and generalization capabilities across diverse environmental conditions. A proposed experiment could involve developing a comparative framework that tests these hybrid models using real-world data collected from different locations (urban, industrial, and rural), evaluating their predictive accuracy, adaptability to local trends, and computational efficiency on embedded platforms.
- Integrating multimodal data sources, such as satellite imagery, meteorological parameters, and vehicle emissions data, can enhance model accuracy and provide deeper insights into air quality dynamics. A potential idea would involve building a data fusion pipeline that combines ground-level sensor data with open-access satellite imagery and meteorological APIs. This pipeline could be tested using AI models to evaluate the benefits of multimodal integration in predictive accuracy and spatial resolution.
- Furthermore, the reliability of AI models must be tested with live data, both on the server and on the edge, to confirm their efficacy in real-world scenarios. To address this, we propose deploying a testbed of edge devices running lightweight AI models for imputation. These devices could be installed in key environmental zones and compared against centralized server-based models to evaluate latency, accuracy, and energy consumption, using a set of key performance indicators. Such a deployment would validate model performance and communication reliability using protocols like WiFi, LoRa, or NB-IoT. Advances in edge computing and low-power AI hardware offer new real-time air quality monitoring opportunities in remote and underserved areas.
- While AI and IoT-based air quality monitoring systems have shown great promise in addressing environmental challenges, continued interdisciplinary research efforts and collaboration between academia, industry, and policymakers are key to overcoming existing limitations and driving widespread adoption of these technologies for global air quality management and public health improvement.

9 Glossary

- **Air Quality Index (AQI)**: A standardized index providing reports on daily air quality levels and associated health implications based on pollutant concentrations.
- **Air Quality Index (AQI)**: A numerical scale employed to report daily air quality levels

and associated health risks based on pollutant concentrations.

- **Anomaly Detection**: The systematic process of identifying abnormal or unexpected patterns in data that may indicate sensor malfunction or rare events.
- **Artificial Intelligence (AI)**: The emulation of human intelligence processes by machines, particularly computer-based systems.
- **Autoencoder**: A neural network architecture designed to learn efficient data representations by compressing and reconstructing inputs, frequently employed for noise reduction and anomaly identification.
- **Convolutional Neural Network (CNN)**: A class of deep neural networks well-suited for capturing spatial dependencies in structured input such as images.
- **Data Imputation**: The systematic process of estimating and filling missing or corrupted data values within a dataset.
- **Deep Learning (DL)**: A class of ML algorithms utilizing multiple layers of neural networks to model complex data relationships.
- **Edge Computing**: Data processing performed proximal to the data source, reducing latency and minimizing reliance on cloud infrastructure.
- **Exposure Evaluation**: Estimation of human exposure to air pollutants based on environmental monitoring and behavioral patterns.
- **Feature Extraction**: The transformation of raw data into structured features that enhance machine learning model performance.
- **Generative Adversarial Network (GAN)**: A DL model in which a generator and a discriminator operate adversarially, frequently employed for synthetic data generation or missing data reconstruction.
- **GSM / GPRS / BLE / Zigbee / Xbee**: Various wireless communication protocols used in IoT, featuring varying transmission ranges, energy requirements, and data rates.
- **Internet of Things (IoT)**: A network of physical devices embedded with sensors and software, enabling data exchange via the internet.
- **LoRa / LoRaWAN**: Wireless protocols ideal for long-range, low-power communication within Internet of Things ecosystems.
- **Long Short-Term Memory (LSTM)**: A specific class of RNNs capable of learning long-term dependencies in sequential datasets.
- **Low-Cost Sensors (LCS)**: Affordable environmental sensing units, generally less precise than reference-grade counterparts, yet suitable for large-scale IoT deployments.
- **MAE (Mean Absolute Error)**: The average of absolute deviations between predicted and observed values.
- **Machine Learning (ML)**: A subfield of AI enabling systems to learn from data and improve performance in the absence of explicit programming.
- **MAPE (Mean Absolute Percentage Error)**: The average of absolute percentage deviations between predictions and ground-truth values.
- **Microcontroller**: Small-scale computing units employed in embedded systems to control sensor interfaces and process data.
- **Overfitting / Underfitting**: Modeling issues where the model captures excessive noise or insufficient structure from training data, impairing generalization.
- **Particulate Matter (PM2.5, PM10)**: Airborne particles with aerodynamic diameters smaller than 2.5μm or 10μm, respectively, associated with adverse health effects.
- **Precision / Recall / F1 Score**: Classification metrics: precision quantifies exactness,

recall evaluates completeness, and the F1 score balances the two.

- **Reconstruction Error**: The difference between original and reconstructed data in autoencoders, used to identify significant deviations or anomalies.
- **Recurrent Neural Network (RNN)**: A neural network architecture capable of processing sequential data through feedback connections that retain contextual information.
- **RMSE (Root Mean Square Error)**: The square root of the mean of squared deviations, placing greater emphasis on larger errors.
- R^2**(Coefficient of Determination)**: A statistical metric indicating the extent to which predicted values approximate observed values, with values closer to 1 indicating better performance.
- **Sensor Calibration**: The alignment of sensor outputs with a reference standard to maintain measurement accuracy.
- **Sensor Drift**: Progressive deviation in sensor output over time due to environmental exposure or hardware degradation.
- **Sensor Drift**: The progressive shift in sensor response over time attributable to aging, environmental conditions, or hardware instability.
- **Short-term Forecasting**: Predictive modeling of air quality conditions over short temporal horizons using a combination of historical and real-time observations.
- **Spatio-temporal Data**: Data incorporating both spatial and temporal dimensions, commonly encountered in environmental monitoring applications.
- **Supervised Learning**: An ML approach in which models are trained on labeled datasets to perform predictive tasks.
- **Unsupervised Learning**: An ML technique employed to uncover hidden data patterns without labeled outputs, such as clustering.
- **Volatile Organic Compounds (VOCs)**: Organic substances that readily vaporize at ambient conditions, frequently linked to adverse indoor and outdoor air quality.

Acknowledgements The authors acknowledge the support of the National Secretariat of Science, Technology, and Innovation (SENACYT) of Panama under Grant No. 157-2023 FID23-078 and the National Research System (SNI) of Panama. The authors also thank the Universidad Tecnológica de Panamá and the Centro de Estudios Multidisciplinarios en Ciencias, Ingeniería y Tecnología AIP (CEMCIT AIP) for their support during this research project.

Author contributions Antony García: Conceptualization; Methodology; Investigation; Writing - Original Draft. Yessica Sáez: Writing - Review & Editing; Visualization. Itamar Harris: Writing - Original Draft; Investigation. Xinming Huang: Formal analysis; Supervision. Edwin Collado: Writing - Review & Editing; Formal analysis; Funding acquisition.

Funding Edwin Collado reports financial support was provided by the National Secretariat of Science, Technology, and Innovation (SENACYT) of Panama. Both Edwin Collado and Yessica Saez report financial support provided by National Research System (SNI) of Panama.

Data availability No datasets were generated or analysed during the current study.

Materials availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare that no Conflict of interest, financial or otherwise, are associated with this work.

Ethical approval This paper does not contain any studies with human participants or animals performed by any of the authors.

References

Abimannan S, El-Alfy E-SM, Hussain S, Chang Y-S, Shukla S, Satheesh D, Breslin JG (2023) Towards federated learning and multi-access edge computing for air quality monitoring: literature review and assessment. Sustainability 15(18):13951. https://doi.org/10.3390/su151813951

Agbo B, Al-Aqrabi H, Hill R, Alsboui T (2022) Missing data imputation in the internet of things sensor networks. Future Internet 14(5):143. https://doi.org/10.3390/fi14050143

AI-Thinker (2020) ESP-01 wi-fi module - datasheet and specifications. https://wiki.ai-thinker.com/esp8266:esp-01

Al-Ali AR, Al-Ali A-R, Zualkernan IA, Zualkernan IA, Aloul F, Aloul FA (2010) A mobile GPRS-sensors array for air pollution monitoring. IEEE Sens J. https://doi.org/10.1109/jsen.2010.2045890

Ali S, Glass T, Parr B, Potgieter J, Alam F (2021) Low cost sensor with IoT LoRaWAN connectivity and machine learning-based calibration for air pollution monitoring. IEEE Trans Instrum Meas 70:1–11. https://doi.org/10.1109/TIM.2020.3034109

Alkabbani H, Ramadan A, Zhu Q, Elkamel A (2022) An improved air quality index machine learning-based forecasting with multivariate data imputation approach. Atmosphere 13(7):1144. https://doi.org/10.3390/atmos13071144

Arduino Inc. (2020) Arduino yún board - technical specifications. https://docs.arduino.cc/retired/boards/arduino-yun/

Badura M, Batog P, Drzeniecka-Osiadacz A, Modzel P (2018) Evaluation of low-cost sensors for ambient PM2.5 monitoring. J Sens 2018(1):5096540. https://doi.org/10.1155/2018/5096540

Baller SP, Jindal A, Chadha M, Gerndt M (2021) DeepEdgeBench: Benchmarking deep neural networks on edge devices. In: 2021 IEEE international conference on cloud engineering (IC2E), pp. 20–30. https://doi.org/10.1109/IC2E52221.2021.00016. https://ieeexplore.ieee.org/abstract/document/9610432?casa_token=zOG7J2eGNz8AAAAA:FYs6TFT5Q4mgKF56xEO7zbxCagKgrPTy_Py64XEBUp6aYqss_HPeq0GguRYX5RpqJtIu8CypYg Accessed 2025-02-08

Benammar M, Benammar M, Abdaoui A, Abdaoui A, Ahmad SHM, Ahmad SHM, Touati F, Touati F, Kadri A, Kadri A (2018) A modular IoT platform for real-time indoor air quality monitoring. Sensors. https://doi.org/10.3390/s18020581

Bernal D, Vanegas A, Pachon Arttesano J (2020) CanAirIO. CanAirIO

Bharathi PD, Narayanan VA, Sivakumar PB (2022) Fog computing enabled air quality monitoring and prediction leveraging deep learning in IoT. J Intell Fuzzy Syst 43(5):5621–5642

Bílek J, Bílek O, Maršolek P, Buček P (2021) Ambient air quality measurement with low-cost optical and electrochemical sensors: an evaluation of continuous year-long operation. Environments 8(11):114. https://doi.org/10.3390/environments8110114

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Brown A, Franken P, Bonner S, Dolezal N, Moross J (2016) Safecast: successful citizen-science for radiation measurement and communication after Fukushima. J Radiol Prot 36(2):82. https://doi.org/10.1088/0952-4746/36/2/S82

Buonanno G, Stabile L, Morawska L, Giovinco G, Querol X (2017) Do air quality targets really represent safe limits for lung cancer risk? Sci Total Environ 580:74–82. https://doi.org/10.1016/j.scitotenv.2016.11.216

Bush T, Papaioannou N, Leach F, Pope FD, Singh A, Thomas GN, Stacey B, Bartington S (2022) Machine learning techniques to improve the field performance of low-cost air quality sensors. Atmos Meas Tech 15(10):3261–3278. https://doi.org/10.5194/amt-15-3261-2022

Castell N, Dauge FR, Schneider P, Vogt M, Lerner U, Fishbain B, Broday D, Bartonova A (2017) Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environ Int 99:293–302. https://doi.org/10.1016/j.envint.2016.12.007

Centenaro M, Vangelista L, Zanella A, Zorzi M (2016) Long-range communications in unlicensed bands: the rising stars in the IoT and smart city scenarios. IEEE Wirel Commun 23(5):60–67. https://doi.org/10.1109/MWC.2016.7721743

Chen L-J, Ho Y-H, Lee H-C, Wu H-C, Liu H-M, Hsieh H-H, Huang Y-T, Lung S-CC (2017) An open framework for participatory PM2.5 monitoring in smart cities. IEEE Access: Pract Innov, Open Solut 5:14441–14454. https://doi.org/10.1109/ACCESS.2017.2723919

Chen Y, Chen X, Xu A, Sun Q, Peng X (2022) A hybrid CNN-transformer model for ozone concentration prediction. Air Qual, Atmosp Health 15(9):1533–1546. https://doi.org/10.1007/s11869-022-01197-w

Chen M, Zhu H, Chen Y, Wang Y (2022) A novel missing data imputation approach for time series air quality data based on logistic regression. Atmosphere 13(7):1044. https://doi.org/10.3390/atmos13071044

Chung W-Y, Chung W-Y, Oh S-J, Oh S-J (2006) Remote monitoring system with wireless sensors module for room environment. Sens Actuators B-Chem. https://doi.org/10.1016/j.snb.2005.02.023

Collado E, Calderón S, Cedeño B, León OD, Centella M, García A, Sáez Y (2024) Open-source internet of things (IoT)-based air pollution monitoring system with protective case for tropical environments. HardwareX 19:00560. https://doi.org/10.1016/j.ohx.2024.e00560

Cui B, Liu M, Li S, Jin Z, Zeng Y, Lin X (2023) Deep learning methods for atmospheric PM2.5 prediction: a comparative study of transformer and CNN-LSTM-attention. Atmos Pollut Res 14(9):101833. https://doi.org/10.1016/j.apr.2023.101833

Das P, Ghosh S, Chatterjee S, De S (2022) A Low cost outdoor air pollution monitoring device with power controlled built-in PM sensor. IEEE Sens J 22(13):13682–13695. https://doi.org/10.1109/JSEN.2022.3175821

David R, Duke J, Jain A, Janapa Reddi V, Jeffries N, Li J, Kreeger N, Nappier I, Natraj M, Wang T, Warden P, Rhodes R (2021) TensorFlow lite micro: embedded machine learning for TinyML systems. Proc Mach Learn Syst 3:800–811

Dhingra S, Dhingra S, Babu MR, Madda RB, Madda RB, Gandomi AH, Gandomi AH, Patan R, Patan R, Daneshmand M, Daneshmand M (2019) Internet of things mobile-air pollution monitoring system (IoT-Mobair). IEEE Internet Things J. https://doi.org/10.1109/jiot.2019.2903821

Dinculeană D, Cheng X (2019) Vulnerabilities and limitations of MQTT protocol used between IoT devices. Appl Sci 9(5):848. https://doi.org/10.3390/app9050848

Du S, Li T, Yang Y, Horng S-J (2021) Deep air quality forecasting using hybrid deep learning framework. IEEE Trans Knowl Data Eng 33(6):2412–2424. https://doi.org/10.1109/TKDE.2019.2954510

Dutta DL, Bharali S (2021) TinyML meets IoT: a comprehensive survey. Internet Things 16:100461. https://doi.org/10.1016/j.iot.2021.100461

Espressif Systems (2020) ESP8266EX datasheet. https://www.espressif.com/sites/default/files/documentation/0a-esp8266ex_datasheet_en.pdf

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

Ganesh PSSP, Venkataraman H (2021) RF-based wireless communication for shallow water networks: survey and analysis. Wirel Pers Commun 120(4):3415–3441. https://doi.org/10.1007/s11277-021-09068-w

Garbagna L, Saheer LB, Dar Oghaz MM (2025) AI-driven approaches for air pollution modelling: a comprehensive systematic review. Environ Pollut 373:125937. https://doi.org/10.1016/j.envpol.2025.125937

Gómez-Carracedo MP, Andrade JM, López-Mahía P, Muniategui S, Prada D (2014) A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. Chemom Intell Lab Syst 134:23–33. https://doi.org/10.1016/j.chemolab.2014.02.007

Haque KF, Abdelgawad A, Yelamarthi K (2022) Comprehensive performance analysis of Zigbee communication: an experimental approach with XBee S2C module. Sensors 22(9):3245. https://doi.org/10.3390/s22093245

Hasenkopf CA, Flasher J, Veerman O, DeWitt HL (2015) OpenAQ: a platform to aggregate and freely share global air quality data. In: AGU fall meeting abstracts, vol. 2015, pp. 31–0097

Hassan NA, Hashim Z, Hashim JH (2016) Impact of climate change on air quality and public health in urban areas. Asia Pac J Public Health 28(2–suppl):38–48. https://doi.org/10.1177/1010539515592951
Hill DJ, Minsker BS (2010) Anomaly detection in streaming environmental sensor data: a data-driven modeling approach. Environ Model Softw 25(9):1014–1022. https://doi.org/10.1016/j.envsoft.2009.08.010
Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
Hossain E, Shariff MAU, Hossain MS, Andersson K (2021) A novel deep learning approach to predict air quality index. In: Kaiser MS, Bandyopadhyay A, Mahmud M, Ray K (eds) Proceedings of international conference on trends in computational and cognitive engineering. Springer, Singapore, pp 367–381
Hoyos CD, Herrera-Mejía L, Roldán-Henao N, Isaza A (2019) Effects of fireworks on particulate matter concentration in a narrow valley: the case of the Medellín metropolitan area. Environ Monit Assess 192(1):6. https://doi.org/10.1007/s10661-019-7838-9
Hůnová I, Hůnová I, Šantroch J, Šantroch J, Ostatnická J, Ostatnická J (2004) Ambient air quality and deposition trends at rural stations in the Czech Republic during 1993–2001. Atmos Environ. https://doi.org/10.1016/j.atmosenv.2003.10.032
Husein NAA, Husein NAA, Rahman AHA, Rahman AHA, Dahnil DP, Dahnil DP, Dahnil DP (2019) Evaluation of LoRa-based air pollution monitoring system. Int J Adv Comput Sci Appl. https://doi.org/10.14569/ijacsa.2019.0100753
Idrees Z, Zou Z, Zheng L (2018) Edge computing based IoT architecture for low cost air pollution monitoring systems: a comprehensive system analysis, design considerations & development. Sensors 18(9):3021. https://doi.org/10.3390/s18093021
Idroes GM, Noviandy TR, Maulana A, Zahriah Z, Suhendrayatna S, Suhartono E, Khairan K, Kusumo F, Helwani Z, Abd Rahman S (2023) Urban air quality classification using machine learning approach to enhance environmental monitoring. Leuser J Environ Stud 1(2):62–68. https://doi.org/10.60084/ljes.v1i2.99
Jabbar WA, Jabbar WA, Subramaniam T, Subramaniam T, Ong AE, Ong AE, Shu'Ib MI, Shu'Ib MI, Wu W, Wu W, Oliveira MAd, Oliveira MAd (2022) LoRaWAN-based IoT system implementation for long-range outdoor air quality monitoring. Internet Things 19:100540. https://doi.org/10.1016/j.iot.2022.100540
Janarthanan R, Partheeban P, Somasundaram K, Elamparithi PN (2021) A deep learning approach for prediction of air quality index in a metropolitan city. Sustain Cities Soc 67:102720. https://doi.org/10.1016/j.scs.2021.102720
Jesus G, Casimiro A, Oliveira A (2021) Using machine learning for dependable outlier detection in environmental monitoring systems. ACM Trans Cyber-Phys Syst 5(3):29–12930. https://doi.org/10.1145/3445812
Junger WL, Leon APd (2015) Imputation of missing data in time series for air pollutants. Atmos Environ 102:96–104. https://doi.org/10.1016/j.atmosenv.2014.11.049
Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M (2004) Methods for imputation of missing values in air quality data sets. Atmos Environ 38(18):2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026
Kaginalkar A, Kumar S, Gargava P, Niyogi D (2021) Review of urban computing in air quality management as smart city service: an integrated IoT, AI, and cloud technology perspective. Urban Climate 39:100972. https://doi.org/10.1016/j.uclim.2021.100972
Kelly FJ, Fussell JC (2015) Air pollution and public health: emerging hazards and improved understanding of risk. Environ Geochem Health 37(4):631–649. https://doi.org/10.1007/s10653-015-9720-1
Kim T, Kim J, Yang W, Lee H, Choo J (2021) Missing value imputation of time-series air-quality data via deep neural networks. Int J Environ Res Public Health 18(22):12213. https://doi.org/10.3390/ijerph182212213
Kim D, Han H, Wang W, Kang Y, Lee H, Kim HS (2022) Application of deep learning models and network method for comprehensive air-quality index prediction. Appl Sci 12(13):6699. https://doi.org/10.3390/app12136699
Kok I, Şimşek MU, Özdemir S (2017) A deep learning model for air quality prediction in smart cities. In: 2017 IEEE international conference on big data (Big Data), pp. 1983–1990. https://doi.org/10.1109/BigData.2017.8258144
Kosmidis E, Syropoulou P, Tekes S, Schneider P, Spyromitros-Xioufis E, Riga M, Charitidis P, Moumtzidou A, Papadopoulos S, Vrochidis S, Kompatsiaris I, Stavrakas I, Hloupis G, Loukidis A, Kourtidis K, Georgoulias AK, Alexandri G (2018) hackAIR: towards raising awareness about air quality in europe by developing a collective online platform. ISPRS Int J Geo Inf 7(5):187. https://doi.org/10.3390/ijgi7050187
Kotsiantis SB (2011) Decision trees: a recent overview. Artif Intell Rev 39(4):261–283. https://doi.org/10.1007/s10462-011-9272-4

Kravchenko J, Akushevich I, Abernethy AP, Holman S, Ross WG Jr, Lyerly HK (2014) Long-term dynamics of death rates of emphysema, asthma, and pneumonia and improving air quality. Int J Chron Obstruct Pulmon Dis 9:613–627. https://doi.org/10.2147/COPD.S59995

Krogh A (2008) What are artificial neural networks? Nat Biotechnol 26(2):195–197. https://doi.org/10.1038/nbt1386

Kularatna N, Kularatna N, Sudantha BH, Sudantha BH (2008) An environmental air pollution monitoring system based on the IEEE 1451 standard for low cost requirements. IEEE Sens J. https://doi.org/10.1109/jsen.2008.917477

Kumar S, Jasuja A (2017) Air quality monitoring system based on IoT using Raspberry Pi. In: 2017 International conference on computing, communication and automation (ICCCA), pp. 1341–1346. https://doi.org/10.1109/CCAA.2017.8230005. https://ieeexplore.ieee.org/document/8230005. Accessed 12 Feb 2025

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. Neural Comput 1(4):541–551. https://doi.org/10.1162/neco.1989.1.4.541

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539

Li X, Peng L, Hu Y, Shao J, Chi T (2016) Deep learning architecture for air quality predictions. Environ Sci Pollut Res 23(22):22408–22417. https://doi.org/10.1007/s11356-016-7812-9

Li Z, Zhang X, Dong Z (2022) TSF-transformer: a time series forecasting model for exhaust gas emission using transformer. Appl Intell 53(13):17211–17225. https://doi.org/10.1007/s10489-022-04326-1

Liang Y, Xia Y, Ke S, Wang Y, Wen Q, Zhang J, Zheng Y, Zimmermann R (2023) AirFormer: predicting nationwide air quality in China with transformers. Proc AAAI Conf Artif Intell 37(12):14329–14337. https://doi.org/10.1609/aaai.v37i12.26676

Liu H, Li Q, Yu D, Gu Y (2019) Air quality index and air pollutant concentration prediction based on machine learning algorithms. Appl Sci 9(19):4069. https://doi.org/10.3390/app9194069

Lonzetta AM, Cope P, Campbell J, Mohd BJ, Hayajneh T (2018) Security vulnerabilities in Bluetooth technology as used in IoT. J Sens Actuator Netw 7(3):28. https://doi.org/10.3390/jsan7030028

Ma Q, Lee W-C, Fu T-Y, Gu Y, Yu G (2020) MIDIA: exploring denoising autoencoders for missing data imputation. Data Min Knowl Disc 34(6):1859–1897. https://doi.org/10.1007/s10618-020-00706-8

Maag B, Zhou Z, Thiele L (2018) A survey on sensor calibration in air pollution monitoring deployments. IEEE Internet Things J 5(6):4857–4870. https://doi.org/10.1109/JIOT.2018.2853660

Mahajan S, Gabrys J, Armitage J (2021) AirKit: A citizen-sensing toolkit for monitoring air quality. Sensors 21(12):4044. https://doi.org/10.3390/s21124044

Mahajan S, Chung M-K, Martinez J, Olaya Y, Helbing D, Chen L-J (2022) Translating citizen-generated air quality data into evidence for shaping policy. Humanit Soc Sci Commun 9(1):122. https://doi.org/10.1057/s41599-022-01135-2

Mahajan M, Kumar S, Pant B, Tiwari UK (2020) Incremental outlier detection in air quality data using statistical methods. In: 2020 International conference on data analytics for business and industry: way towards a sustainable economy (ICDABI), pp. 1–5. https://doi.org/10.1109/ICDABI51230.2020.9325683. https://ieeexplore.ieee.org/abstract/document/9325683. Accessed 05 Sep 2024

Maltare NN, Vahora S (2023) Air quality index prediction using machine learning for Ahmedabad city. Digital Chem Eng 7:100093. https://doi.org/10.1016/j.dche.2023.100093

Manpreet, Malhotra J (2015) ZigBee technology: current status and future scope. In: 2015 International conference on computer and computational sciences (ICCCS), pp. 163–169. https://doi.org/10.1109/ICCACS.2015.7361343. https://ieeexplore.ieee.org/abstract/document/7361343. Accessed 06 April 2025

Marinov MB, Topalov I, Gieva E, Nikolov G (2016) Air quality monitoring in urban environments. In: 2016 39th International spring seminar on electronics technology (ISSE), pp. 443–448. https://doi.org/10.1109/ISSE.2016.7563237. https://ieeexplore.ieee.org/document/7563237. Accessed 25 July 2024

Marques G, Pitarma R (2019) A cost-effective air q1uality supervision solution for enhanced living environments through the internet of things. Electronics 8(2):170. https://doi.org/10.3390/electronics8020170

Martínez J, Saavedra A, Garcia-Nieto PJ, Piñeiro JI, Iglesias C, Taboada J, Sancho J, Pastor J (2014) Air quality parameters outliers detection using functional data analysis in the Langreo urban area (Northern Spain). Appl Math Comput 241:1–10. https://doi.org/10.1016/j.amc.2014.05.004

Martinez B, Adelantado F, Bartoli A, Vilajosana X (2019) Exploring the performance boundaries of NB-IoT. IEEE Internet Things J 6(3):5702–5712. https://doi.org/10.1109/JIOT.2019.2904799

Martín-Garín A, Martín-Garín A, Millán-García JA, Millán-García JA, Bairi A, Bairi A, Bairi A, Millán-Medel J, Millán-Medel J, Sala-Lizarraga JM, Sala-Lizarraga JM (2018) Environmental monitoring system based on an open source platform and the internet of things for a building energy retrofit. Autom Constr. https://doi.org/10.1016/j.autcon.2017.12.017

Marzouk M, Atef M (2022) Assessment of indoor air quality in academic buildings using IoT and deep learning. Sustainability 14(12):7015. https://doi.org/10.3390/su14127015

Ma Y, Yang S, Huang Z, Hou Y, Cui L, Yang D (2014) Hierarchical air quality monitoring system design. In: 2014 International symposium on integrated circuits (ISIC), pp. 284–287. https://doi.org/10.1109/ISICIR.2014.7029544. https://ieeexplore.ieee.org/document/7029544. Accessed 12 Feb 2025

Microchip Technology Inc. (2016) ATmega1281/1280 AVR microcontroller datasheet. https://ww1.microchip.com/downloads/en/DeviceDoc/Atmel-2549-8-bit-AVR-Microcontroller-ATmega640-1280-1281-2560-2561_datasheet.pdf

Microchip Technology Inc. (2016) ATmega2560/1280 AVR microcontroller datasheet. https://ww1.microchip.com/downloads/en/DeviceDoc/Atmel-2549-8-bit-AVR-Microcontroller-ATmega640-1280-1281-2560-2561_datasheet.pdf

Microchip Technology Inc. (2016) PIC32MM family data sheet. https://ww1.microchip.com/downloads/en/DeviceDoc/PIC32MM-Data-Sheet-60001320C.pdf

Mishra M, Chen P-H, Bisquera W, Lin G-Y, Le T-C, Dejchanchaiwong R, Tekasakul P, Jhang C-W, Wu C-J, Tsai C-J (2023) Source-apportionment and spatial distribution analysis of VOCs and their role in ozone formation using machine learning in central-west Taiwan. Environ Res 232:116329. https://doi.org/10.1016/j.envres.2023.116329

Mitreska Jovanovska E, Batz V, Lameski P, Zdravevski E, Herzog MA, Trajkovik V (2023) Methods for urban air pollution measurement and forecasting: challenges, opportunities, and solutions. Atmosphere 14(9):1441. https://doi.org/10.3390/atmos14091441

...Morawska L, Thai PK, Liu X, Asumadu-Sakyi A, Ayoko G, Bartonova A, Bedini A, Chai F, Christensen B, Dunbabin M, Gao J, Hagler GSW, Jayaratne R, Kumar P, Lau AKH, Louie PKK, Mazaheri M, Ning Z, Motta N, Mullins B, Rahman MM, Ristovski Z, Shafiei M, Tjondronegoro D, Westerdahl D, Williams R (2018) Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: how far have they gone? Environ Int 116:286–299. https://doi.org/10.1016/j.envint.2018.04.018

Mukaro R, Mukaro R, Carelse XF, Carelse XF, Carelse XF (1997) A serial communication program for accessing a microcontroller-based data-acquisition system. Comput Geosci. https://doi.org/10.1016/s0098-3004(97)00065-4

Mukaro R, Mukaro R, Carelse XF, Carelse XF (1999) A microcontroller-based data acquisition system for solar radiation and environmental monitoring. IEEE Trans Instrum Meas. https://doi.org/10.1109/19.816142

Navares R, Aznarte JL (2020) Predicting air quality with deep learning LSTM: towards comprehensive models. Eco Inform 55:101019. https://doi.org/10.1016/j.ecoinf.2019.101019

Nikzad N, Verma N, Ziftci C, Bales E, Quick N, Zappi P, Patrick K, Dasgupta S, Krueger I, Rosing TS, Griswold WG (2012) CitiSense: improving geospatial environmental assessment of air quality using a wireless personal exposure monitoring system. In: Proceedings of the conference on wireless health. Wh '12. Association for computing Machinery, San Diego, California and New York, NY, USA. https://doi.org/10.1145/2448096.2448107

Noble WS (2006) What is a support vector machine? Nat Biotechnol 24(12):1565–1567. https://doi.org/10.1038/nbt1206-1565

open-seneca. open-seneca| Air Quality Sensing| Pollution Mapping. https://www.open-seneca.org/ Accessed 07 April 2025

Osman E, Banerjee C, Poonia AS (2024) HDLP: air quality modeling with hybrid deep learning approaches and particle swarm optimization. Innov Syst Softw Eng 20(3):287–299. https://doi.org/10.1007/s11334-024-00559-0

Ottosen T-B, Kumar P (2019) Outlier detection and gap filling methodologies for low-cost air quality measurements. Environ Sci Process Impacts 21(4):701–713. https://doi.org/10.1039/C8EM00593A

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. https://doi.org/10.1136/bmj.n71

Palomeque-Mangut S, Palomeque-Mangut S, Meléndez F, Meléndez F, Gómez-Suárez J, Gómez-Suárez J, Frutos-Puerto S, Frutos-Puerto S, Arroyo P, Arroyo P, Pinilla-Gil E, Gil EP, Lozano J, Lozano J (2022) Wearable system for outdoor air quality monitoring in a WSN with cloud computing: design, validation and deployment. Chemosphere. https://doi.org/10.1016/j.chemosphere.2022.135948

Park D, Yoo G-W, Park S-H, Lee J-H (2021) Assessment and calibration of a low-cost PM2.5 sensor using machine learning (HybridLSTM neural network): feasibility study to build an air quality monitoring system. Atmosphere 12(10):1306. https://doi.org/10.3390/atmos12101306

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):1–24. https://doi.org/10.1186/s40537-023-00746-z

Pereira C, Pinto A, Ferreira D, Aguiar A (2017) Experimental characterization of mobile IoT application latency. IEEE Internet Things J 4(4):1082–1094. https://doi.org/10.1109/JIOT.2017.2689682
Peterson LE (2009) K-nearest neighbor - Scholarpedia. Accessed 07 April 2025
Postolache O, Postolache O, Pereira M, Pereira JMD, Pereira J, Girão PM, Girão PS (2009) Smart sensors network for air quality monitoring applications. IEEE Trans Instrum Meas. https://doi.org/10.1109/tim.2009.2022372
Pradityo F, Surantha N (2019) Indoor air quality monitoring and controlling system based on IoT and fuzzy logic. In: 2019 7th International conference on information and communication technology (ICoICT), pp. 1–6. https://doi.org/10.1109/ICoICT.2019.8835246. https://ieeexplore.ieee.org/document/8835246. Accessed 12 Feb 2025
Purbakawaca R, Yuwono AS, Subrata IDM, Supandi, Alatas H (2022) Ambient air monitoring system with adaptive performance stability. IEEE Access: Pract Innov, Open Solut 10:120086–120105. https://doi.org/10.1109/ACCESS.2022.3222329
Rad AK, Shamshiri RR, Naghipour A, Razmi S-O, Shariati M, Golkar F, Balasundram SK (2022) Machine learning for determining interactions between air pollutants and environmental parameters in three cities of Iran. Sustainability 14(13):8027. https://doi.org/10.3390/su14138027
Rai AC, Kumar P, Pilla F, Skouloudis AN, Sabatino SD, Ratti C, Yasar A, Rickerby D (2017) End-user perspective of low-cost sensors for outdoor air pollution monitoring. Sci Total Environ 607–608:691–705. https://doi.org/10.1016/j.scitotenv.2017.06.266
Raspberry Pi Foundation (2015) Raspberry pi 2 model B. https://www.raspberrypi.com/products/raspberry-pi-2-model-b/
Raspberry Pi Foundation (2016) Raspberry pi 3 model B. https://www.raspberrypi.com/products/raspberry-pi-3-model-b/
Rollo F, Bachechi C, Po L (2023) Anomaly detection and repairing for improving air quality monitoring. Sensors 23(2):640. https://doi.org/10.3390/s23020640
Sai KBK, Sai KBK, Subbareddy SR, Subbareddy SR, Luhach AK, Luhach AK, Luhach AK (2019) IOT based air quality monitoring system using MQ135 and MQ7 with machine learning analysis. Scalable Comput: Pract Exp. https://doi.org/10.12694/scpe.v20i4.1561
Samal KKR, Babu KS, Das SK (2021) Temporal convolutional denoising autoencoder network for air pollution prediction with missing values. Urban Climate 38:100872. https://doi.org/10.1016/j.uclim.2021.100872
Samie F, Bauer L, Henkel J (2019) From cloud down to things: an overview of machine learning in internet of things. IEEE Internet Things J 6(3):4921–4934. https://doi.org/10.1109/JIOT.2019.2893866
Samuel SSI (2016) A review of connectivity challenges in IoT-smart home. In: 2016 3rd MEC international conference on big data and smart city (ICBDSC), pp. 1–4. https://doi.org/10.1109/ICBDSC.2016.7460395. https://ieeexplore.ieee.org/abstract/document/7460395. Accessed 06 April 2025
Sarnat JA, Holguin F (2007) Asthma and air quality. Curr Opin Pulm Med 13(1):63–66. https://doi.org/10.1097/MCP.0b013e3280117d25
Schubert E, Sander J, Ester M, Kriegel HP, Xu X (2017) DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Trans Database Syst 42(3):1–21. https://doi.org/10.1145/3068335
Sewor C, Obeng AA, Amegah AK (2021) The Ghana urban air quality project (Ghair): bridging air pollution data gaps in Ghana. Clean Air J 31(1):1–2. https://doi.org/10.17159/caj/2021/31/1.11172
Shaadan N, Rahim NAM (2019) Imputation analysis for time series air quality (PM10) data set: a comparison of several methods. J Phys: Conf Ser 1366(1):012107. https://doi.org/10.1088/1742-6596/1366/1/012107
Sparks L, Sumner G (1984) A Microcomputer-based weather station monitoring system. J Microcomput Appl 7(3):233–242. https://doi.org/10.1016/0745-7138(84)90054-X
STMicroelectronics (2021) STM32F103x8/xB datasheet. https://www.st.com/resource/en/datasheet/stm32f103c8.pdf
STMicroelectronics (2021) STM32L476RG datasheet. https://www.st.com/resource/en/datasheet/stm32l476rg.pdf
Sunyer J, Jarvis D, Gotschi T, Garcia-Esteban R, Jacquemin B, Aguilera I, Ackerman U, De Marco R, Forsberg B, Gislason T, Heinrich J, Norbäck D, Villani S, Künzli N (2006) Chronic bronchitis and urban air pollution in an international study. Occup Environ Med 63(12):836–843. https://doi.org/10.1136/oem.2006.027995
Tancev G, Toro FG (2022) Variational Bayesian calibration of low-cost gas sensor systems in air quality monitoring. Meas: Sens 19:100365. https://doi.org/10.1016/j.measen.2021.100365
Tosi J, Taffoni F, Santacatterina M, Sannino R, Formica D (2017) Performance evaluation of bluetooth low energy: a systematic review. Sensors 17(12):2898. https://doi.org/10.3390/s17122898

Tran QA, Dang QH, Le T, Nguyen HT, Le TD (2022) Air quality monitoring and forecasting system using IoT and machine learning techniques. In: 2022 6th International conference on green technology and sustainable development (GTSD), pp. 786–792. https://doi.org/10.1109/GTSD54989.2022.9988756. https://ieeexplore.ieee.org/document/9988756 Accessed 01 Sep 2024

Ullo SL, Sinha GR (2020) Advances in smart environment monitoring systems using IoT and sensors. Sensors 20(11):3113. https://doi.org/10.3390/s20113113

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems, vol 30. Curran Associates Inc, Newry

Vincent P, Larochelle H, Bengio Y, Manzagol P-A (2008) Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on machine learning. ICML '08, pp. 1096–1103. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1390156.1390294. https://dl.acm.org/doi/10.1145/1390156.1390294. Accessed 06 Feb 2025

Wang J, Du P, Hao Y, Ma X, Niu T, Yang W (2020) An innovative hybrid model based on outlier detection and correction algorithm and heuristic intelligent optimization algorithm for daily air quality index forecasting. J Environ Manag 255:109855. https://doi.org/10.1016/j.jenvman.2019.109855

Wang J, Xu W, Zhang Y, Dong J (2022) A novel air quality prediction and early warning system based on combined model of optimal feature extraction and intelligent optimization. Chaos, Solitons Fract 158:112098. https://doi.org/10.1016/j.chaos.2022.112098

Wang Z, Yang Y, Yue S (2022) Air quality classification and measurement based on double output vision transformer. IEEE Internet Things J 9(21):20975–20984. https://doi.org/10.1109/JIOT.2022.3176126

Wardana INK, Gardner JW, Fahmy SA (2022) Estimation of missing air pollutant data using a spatiotemporal convolutional autoencoder. Neural Comput Appl 34(18):16129–16154. https://doi.org/10.1007/s00521-022-07224-2

Wei Y, Jang-Jaccard J, Xu W, Sabrina F, Camtepe S, Boulic M (2023) LSTM-autoencoder-based anomaly detection for indoor air quality time-series data. IEEE Sens J 23(4):3787–3800. https://doi.org/10.1109/JSEN.2022.3230361

Wijesekara WMLKN, Liyanage L (2020) Comparison of imputation methods for missing values in air pollution data: case study on Sydney air quality index. In: Arai K, Kapoor S, Bhatia R (eds) Adv Inf Commun. Springer, Cham, pp 257–269

William P, Paithankar DN, Yawalkar PM, Korde SK, Rajendra A, Pabale, Rakshe DS (2023) Divination of air quality assessment using wnsembling machine learning approach. In: 2023 International conference on artificial intelligence and knowledge discovery in concurrent engineering (ICECONF), pp. 1–10. https://doi.org/10.1109/ICECONF57129.2023.10083751. https://ieeexplore.ieee.org/document/10083751 Accessed 25 July 2024

Wu Z, Ma C, Shi X, Wu L, Dong Y, Stojmenovic M (2022) Imputing missing indoor air quality data with inverse mapping generative adversarial network. Build Environ 215:108896. https://doi.org/10.1016/j.buildenv.2022.108896

Xayasouk T, Lee H, Lee G. Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models. Sustainability 12(6):2570

Xu X, Yoneda M (2021) Multitask air-quality prediction based on LSTM-autoencoder model. IEEE Trans Cybernet 51(5):2577–2586. https://doi.org/10.1109/TCYB.2019.2945999

Xu Y, Zhu Y (2016) When remote sensing data meet ubiquitous urban data: fine-grained air quality inference. In: 2016 IEEE international conference on big data (Big Data), pp. 1252–1261. https://doi.org/10.1109/BigData.2016.7840729. https://ieeexplore.ieee.org/abstract/document/7840729?casa_token=19h3K_aEp1wAAAAA:L0Ac9g2DS3I09GuHaVv7I_wFYc4hq6_uhGdL0Vcmyt_wHxGgaIsgi6xe26J0r0WNN735pdth8Q Accessed 25 July 2024

Yang Y, Li L (2015) A smart sensor system for air quality monitoring and massive data collection. In: 2015 International conference on information and communication technology convergence (ICTC), pp. 147–152. https://doi.org/10.1109/ICTC.2015.7354515

Zhang Z, Zhang S (2023) Modeling air quality PM2.5 forecasting using deep sparse attention-based transformer networks. Int J Environ Sci Technol 20(12):13535–13550. https://doi.org/10.1007/s13762-023-04900-1

Zhao L, Zhao L, Zhao L, Wu W, Wu W, Li S, Li S (2019) Design and implementation of an IoT-Based indoor air quality detector with multiple communication interfaces. IEEE Internet Things J. https://doi.org/10.1109/jiot.2019.2930191

Zheng K, Zhao S, Yang Z, Xiong X, Xiang W (2016) Design and implementation of LPWA-based air quality monitoring system. IEEE Access: Pract Innov, Open Solut 4:3238–3245. https://doi.org/10.1109/ACCESS.2016.2582153

Zhou C, Paffenroth RC (2017) Anomaly detection with robust deep autoencoders. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. KDD '17, pp. 665–674. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3097983.3098052. https://dl.acm.org/doi/10.1145/3097983.3098052. Accessed 06 Feb 2025

Zimmerman N, Presto AA, Kumar SPN, Gu J, Hauryliuk A, Robinson ES, Robinson AL, Subramanian R (2018) A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmos Meas Tech 11(1):291–313. https://doi.org/10.5194/amt-11-291-2018

Authors and Affiliations

Antony Garcia[1,4] · Yessica Saez[1,3] · Itamar Harris[2] · Xinming Huang[4] · Edwin Collado[1,3]

✉ Yessica Saez
yessica.saez@utp.ac.pa

✉ Edwin Collado
edwin.collado@utp.ac.pa

Antony Garcia
antony.garcia@utp.ac.pa

Itamar Harris
itamar.harris@utp.ac.pa

Xinming Huang
xhuang@wpi.edu

1 Facultad de Ingeniería Eléctrica, Universidad Tecnológica de Panamá, Avenida Universidad Tecnológica de Panamá, Panamá 0819-0728, Panama

2 Facultad de Ingeniería Mecánica, Universidad Tecnológica de Panamá, Avenida Universidad Tecnológica de Panamá, Panamá 0819-0728, Panama

3 Centro de Estudios Multidisciplinarios en Ciencias, Ingeniería y Tecnología AIP (CEMCIT AIP), Panamá 0819-0728, Panama

4 Department of Electrical and Computer Engineering, Worcester Polytechnic Institute, Worcester, MA 01609, USA

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH ("Springer Nature").

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users ("Users"), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use ("Terms"). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com