World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Forecasting health outcomes of Air Pollution: A Statistical review of modelling techniques and applications

Pranjali Subhash Sonone * and Abhay Kamlakar Khamborkar

Department of Statistics, Institute of Science, R.T.Road, Civil Lines, Nagpur, Maharashtra, India 440001.

World Journal of Advanced Research and Reviews, 2025, 27(01), 1255-1262

Publication history: Received on 04 June 2025; revised on 12 July 2025; accepted on 14 July 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.27.1.2639

Abstract

This review summarises all recent papers related to the prediction of diseases caused by air pollution. Most of the papers identified through the literature review focused either on health risk prediction alone or air pollutant prediction in various regions. However, very few research articles are based on predictive models that provide health risk predictions and also assess the effects on health. Therefore, the primary objective of this review is to identify models that incorporate both air pollution data and health data to predict diseases. This review encompasses the shift from traditional models to machine learning models in forecasting. This study will be beneficial for future research aimed at identifying diseases caused solely by specific air pollutants.

Keywords: Review of Air Pollution Studies; Machine Learning; Air Pollution; Health Risk; Prediction of Disease

1. Introduction

A study on the impact of air pollution is necessary because it is harming the environment and contributing to millions of deaths worldwide each year. Researchers have also linked air pollutant exposure to severe health problems such as respiratory disease, cardiovascular disease, and premature mortality. The availability of air pollution and health data has provided new opportunities to develop disease prediction models that can estimate the potential health impact of air quality. Recently, researchers have been utilising various machine learning models due to advancements in computational tools. These models can handle high-dimensional data and nonlinear relationships, as well as complex patterns. Therefore, traditional modelling techniques are being replaced with machine learning modelling for time-saving and rapid research analysis. These generated models enhance predictive accuracy and enable real-time forecasting in urban environments, where policy decisions rely on timely data. This review paper summarises the current literature on statistical and machine learning models used to predict health outcomes associated with air pollution. Here, all papers are reviewed based on their methodologies, data sources, strengths and limitations of the study, and the quality of evidence.

2. Materials and Methods

2.1. Search Strategy

A comprehensive literature search was conducted using major academic databases, including Google Scholar, Scopus, PubMed, ResearchGate, Elsevier, and Web of Science. The aim was to identify the original research articles that include air pollution disease prediction modelling using traditional or machine learning techniques to explore the relationship between air pollution exposure and health outcomes. Only recent studies published between 2015 and 2025 were selected. Search queries using the Boolean operators were used.

* Corresponding author: Pranjali Subhash Sonone.

- ("air pollution" OR "particulate matter" OR PM2.5 OR PM10)
- AND ("health effects" OR "respiratory disease" OR "mortality" OR "hospitalisation")
- AND ("modelling" OR "forecasting" OR "statistical model" OR "machine learning" OR "deep learning")

2.2. Study Screening and Selection

Initially, from all database searches, 27,400 records were obtained. We then removed all duplicates and irrelevant articles to ensure a clean dataset. Based on the title, abstract, and full-text article, 980 records were excluded. Following a meticulous screening process, we selected only 94 records, each of which was carefully chosen and reviewed. Most studies using the chosen records are solely concerned with modelling health data or air pollution. Ultimately, we finalised 21 studies that met the criteria for modelling both air pollution data and health data, including disease prediction models. The study selection process is summarised using the PRISMA 2020 flow diagram.

2.3. Risk of Bias Assessment

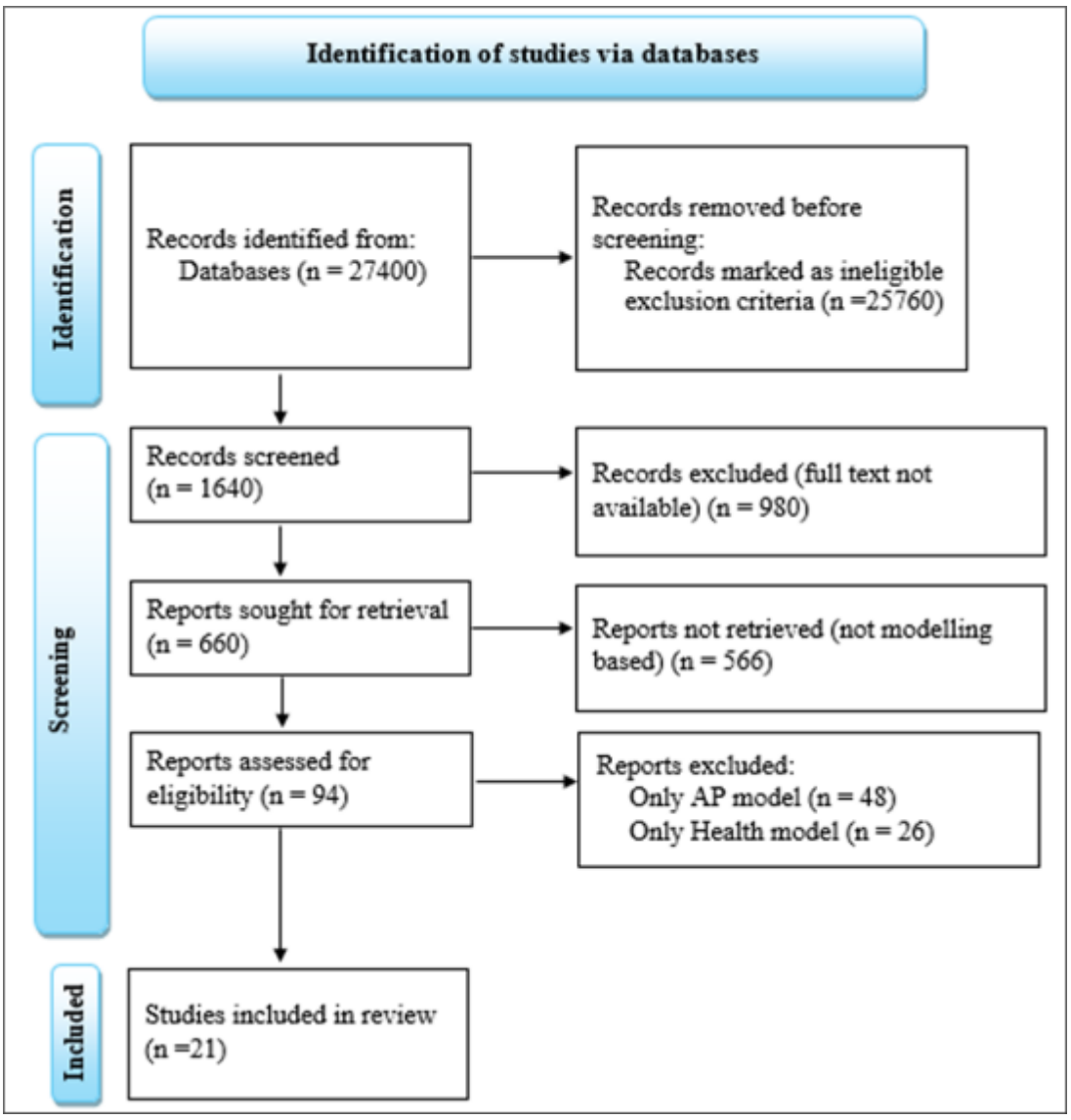

Figure 1 PRISMA flow diagram

This review focused on high-quality studies. Most of these studies clearly defined their objectives, used reliable data sources, and employed the necessary modelling techniques. The risk of bias was assessed using the ROBIS tool. A traffic light plot and summary plot (Figures 2 and 3) indicate that the overall risk is low to moderate. These findings show strong methodological practices in health impact modelling related to air pollution techniques.

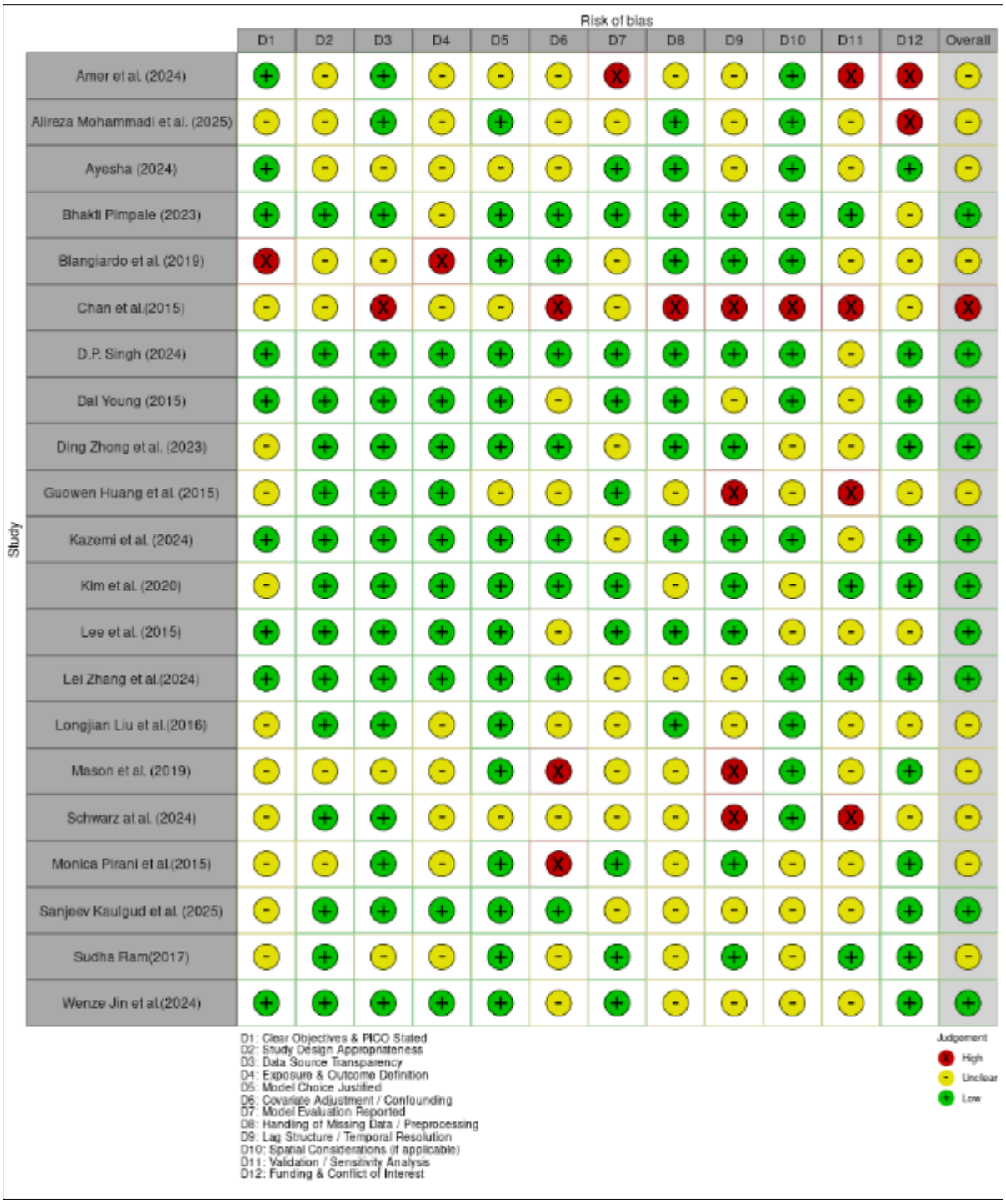

Figure 2 Traffic Light Plot

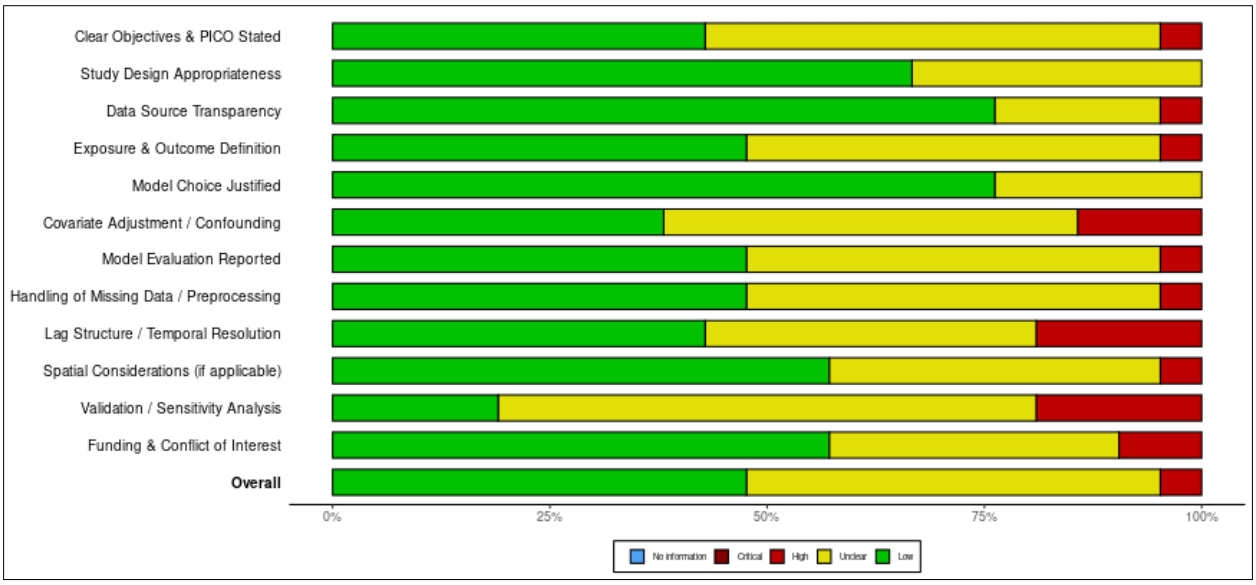

Figure 3 Summary plot

3. Methodology of Included Studies

The selected studies employed various modelling techniques, ranging from traditional to machine learning, depending on the objectives, availability, and complexity of the data. So, these techniques are summarised below.

3.1. Statistical Time-Series and Mixed Models

Most studies rely on regression models because the data are longitudinal and collected from multiple locations. In the study by Kazemi et al. (2024), the authors combined a health impact assessment calculated using the WHO's AirQ+ tool with linear mixed effects (LME) regression to analyse pollutant trends and their health impacts [19]. Hourly data for PM_{10}, NO_2, and O_3 were collected from several cities in Iran, and an LME model was applied, using each city as a random intercept after aggregating pollutant concentrations. This model was chosen to account for repeated monthly measurements within cities and to capture seasonal or nonlinear changes using polynomial terms. Similarly, Zhang et al. (2024) carried out a time series study of daily $PM_{2.5}$ and cardiovascular mortality across 272 cities. [1] For each city, data were first modelled with GAM using quasi-Poisson regression for daily death counts, with adjustments for long-term and seasonal trends (using smooth splines), weather effects, and day-of-week effects. A distributed lag framework was employed to capture same-day and short-delayed $PM_{2.5}$ effects (lag 0-2 days), allowing measurement of short-term pollution spikes related to mortality.

Liu et al. (2016) employed a cross-sectional multilevel regression analysis to investigate the relationship between long-term exposure to fine particulate matter and cardiovascular disease in U.S. cities. The rates of the used derived diseases (stroke, heart disease, diabetes) for 2010-2013 and the average $PM_{2.5}$ concentrations taken from EPA monitors. The nested multilevel model was fitted to reflect geographic clustering and unobserved state effects. Additionally, a hierarchical regression was chosen because it properly partitions variance at the state and county levels, controls for confounding, and provides an unbiased estimate of how a 10 $\mu g/m^3$ increase in long-term $PM_{2.5}$ relates to changes in disease prevalence [2].

3.2. Cohort and Cross-Sectional Designs

Traditional epidemiological designs, often combined with regression models, are also employed in some studies. Jung et al. (2015) surveyed 5443 Korean children to investigate the effect of traffic-related air pollution (TAP) and allergies in their cross-sectional study. GIS is used for TAP exposure measurements, such as the distance from home to major roads (total road length and density within a 200-m radius). Health outcomes (asthma, allergic rhinitis, lung function) were obtained from questionnaires and clinical records. Then, a multivariate logistic regression analysis was used to examine the effects of TAP on health outcomes, given that the outcomes were binary. [3]

A prospective birth cohort study of 736 children was conducted by Hsu et al. (2015) to check the relationship between $PM_{2.5}$ and asthma onset with the use of a high-resolution satellite-based land-use regression model (combining aerosol

optical depth with land-use and meteorological data). Also, a distributed lag models (DLM), which estimate pollutant effects at each week of gestation on asthma, was used to identify the critical exposure window. This approach was used because it can flexibly detect when in pregnancy exposure is most harmful, and also stratified by the child's sex to test interactions.

3.3. Bayesian and Hierarchical Models

A fully Bayesian hierarchical frameworks were used in several studies. Pirani et al. (2015) used this method to assess the relationship between continuous particle exposure and daily respiratory mortality in London. They employed the Dirichlet process mixture model to cluster days based on their multi-pollutant profiles and associated mortality counts. Instead of regressing pollutants separately, they used this method because the high correlation of pollutants can give unstable results. In this process, the number of clusters is a priori detected from the data. Then, days with similar pollutant mixes and mortality form latent clusters, each characterised by a mean exposure vector and mortality risk. Additionally, smooth spline terms for time and temperature are employed to account for seasonal and weather confounding. This approach can handle high-dimensional, correlated exposures and reveal hidden patterns of pollution-related health effects without pre-specifying pollutant combinations[4], [5].

Similarly, Huang (2015) developed a two-stage Bayesian hierarchical model in Scotland. In stage 1, a fusion model was used that combined monitored NO2 data to create fine-scale pollution maps. In stage 2, those estimated NO2 levels were linked to respiratory hospital admissions using a Poisson log-linear regression. This health model has spatial random effects to adjust for residual spatio-temporal autocorrelation. By fusing data sources, the model aimed to improve exposure estimates (especially where monitors are sparse), and the hierarchical approach naturally incorporated uncertainty from pollutant estimation into the disease model [6].

Blangiardo et al. (2019) proposed a fully Bayesian joint model (H2Mjoint) for time-series data in their study, which estimates both "latent" pollutant concentrations and health effects as components of this two-component model. Component 1 models the actual pollutant levels using multivariate autoregressive models, which account for measurement error and correlations among pollutants. Component 2 regresses daily cardiovascular mortality on those latent concentrations, employing Poisson regression with splines for time and weather. Importantly, the joint estimation passes uncertainty from pollutant modelling into the health model, which contrasts with traditional two-step methods that first estimate pollution and then plug it into health regression. By fitting the entire system via MCMC (e.g., in Open BUGS), they can account for high pollutant correlations and missing data, yielding more robust effect estimates [7].

3.4. Machine Learning and Hybrid Predictive Models

Nowadays, due to advancements in studies, machine learning models are used more often. Bhakti Pimpale (2023) in her study used air quality and weather variables to build a multi-output ensemble ML model to forecast daily respiratory outpatient visits. She has used satellite as well as ground-level data, along with historical OPS counts for acute respiratory infections and pneumonia. She has tested 13 algorithms, and the final model was developed by combining Gaussian and extra trees regressors because this model can handle data issues better than any single model. The methodology employed in this study involves extensive preprocessing, including data transformations, PCA/VIF for detecting multicollinearity, stationarity tests, and tuning of 7–8-day lag structures for each target disease. This ML-based approach was used because it can capture complex nonlinear relationships in the data and produce simultaneous forecasts for multiple diseases[9]

Similarly, D.P. Singh (2024) reviewed multiple ML models for predicting lung cancer onset from clinical features. Then, he compared algorithms such as Support Vector Machines, Random Forests, neural networks, and ensemble methods. The electronic health record was used to frame the analysis on these features to classify cancer risk, evaluated by metrics (accuracy, AUC-ROC, etc.). The methodology used for comparing models under a uniform framework, with feature selection and preprocessing (handling missing data, balancing classes) to improve performance. The main objective of this study was to highlight how ML could improve the early detection of lung cancer[10].

Lei Zhang (2024) took a hybrid forecasting approach for an Air Quality Health Index (AQHI) in Guangzhou. They used a combined Random Forest and Adaptive Lasso ("RF-Alasso") to select key pollutants (from $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3) with meteorological factors. Then, they built three time-series models using daily data (2015–2019 for training, 2020–2021 for testing) with LSTM networks. These models were evaluated by metrics (R^2, MAE, RMSE, AIC) on mortality outcomes. The hybrid approach was chosen to leverage the strengths of both classical (ARIMA) and deep learning (LSTM) methods, aiming for more accurate AQHI predictions[1].

In all the above studies, each method requires appropriate data; most studies combined air quality (monitor or satellite) data with health statistics (mortality, hospital admissions, or clinical surveys) and covariates (demographics, weather). So, depending on the type of data model selected, most of the studies show that a hybrid modelled approach is better than selecting a single model. A combination of Machine learning methods with traditional approaches (e.g., Light GBM method with linear regression) yields a better model for assessing health impacts.

4. Results and Discussion

All the reviewed studies collectively provide evidence that both short-term and long-term exposure to air pollutants (especially $PM_{2.5}$, NO_2, and O_3) is strongly associated with adverse health outcomes such as asthma, respiratory tract infections, cardiovascular diseases and premature mortality.

4.1. Health Risks Associated with Specific Pollutants

$PM_{2.5}$ is the most influential pollutant linked with both immediate and delayed health impacts. The short exposure to $PM_{2.5}$ significantly increased cardiovascular mortality [2], and long-term exposure increases the prevalence of stroke, heart disease, and diabetes [11]. It has been found that Nitrogen dioxide (NO_2) associated with vehicular traffic shows a consistent association with respiratory issues and a higher risk of asthma and allergic rhinitis among children living near high-traffic areas [4]. It has also been observed that ozone is associated with acute respiratory symptoms and increased hospital visits [2],[3].

4.2. Temporal and Spatial Dynamics

Time-series studies have demonstrated that increases often follow short-term spikes in pollution levels (especially $PM_{2.5}$ and PM_{10}), resulting in hospital admissions or deaths within 1 to 3 days. Distributed lag models and DLNMs helped reveal these time-sensitive relationships. [6], [12].

Spatial analyses also uncovered important insights. Liu et al. (2016) found regional variations in pollution-related health risks, influenced by demographic and climatic factors. Bayesian hierarchical models [7] were employed to account for such spatial heterogeneity, revealing significant clustering effects that demonstrate how urban form, infrastructure, and socioeconomic variables can influence exposure and vulnerability [2], [6].

4.3. Model Comparisons and Performance Insights

Traditional statistical models, such as Poisson regression, GLM, and GAM, were found to be highly interpretable and suitable for estimating the direct health effects of pollutants. However, their performance may decline when handling high-dimensional data or complex, nonlinear relationships.

Machine learning (ML) models, on the other hand, provided higher predictive accuracy in many cases. For instance, Pimpale (2023) found that an ensemble model combining Gaussian Process and Extra Trees regressors significantly outperformed single algorithms in predicting daily respiratory outpatient visits. The use of ML was especially advantageous when the goal was forecasting rather than causal inference, and when input data included multiple pollutants, weather variables, and temporal lags [9].

Hybrid models, such as ARIMA-LSTM [2], demonstrated superior performance by capturing both linear and non-linear patterns in air quality health index (AQHI) forecasting. These models showed lower error rates (e.g., RMSE) and higher R^2 values, indicating that combining statistical time series forecasting with deep learning components can yield more accurate and adaptive forecasts. However, these models often required large volumes of clean, temporally aligned data and were sensitive to missing values or inconsistent time lags, necessitating robust preprocessing and validation steps [13], [14], [15].

4.4. Role of Bayesian and Explainable Models

Bayesian models provided valuable advantages in uncertainty estimation and integration of multi-level data. The Bayesian profile regression approach by Pirani et al. (2015) enabled the clustering of pollution–mortality profiles without requiring the specification of individual pollutant effects. Blangiardo et al. (2019)'s H2Mjoint model enabled simultaneous estimation of latent pollutant exposures and health outcomes, offering better robustness to measurement error [6], [8].

These approaches, although computationally intensive, are well-suited for complex urban environments where pollutant sources are interdependent and exposure data may be sparse or noisy. They also support interpretability, which is increasingly important in environmental epidemiology.

4.5. Emerging Trends and Future Directions

Several promising trends were observed across the studies: The integration of satellite data and remote sensing into exposure modelling is expanding geographical coverage, especially in low-resource settings. The move from single-pollutant models to multi-pollutant and mixture analysis (e.g., using clustering or joint modelling) reflects a more realistic approach to urban air pollution. Explainable AI techniques, such as feature importance from Random Forest or SHAP values in XGBoost, are being used to interpret the role of individual pollutants in complex models. Natural experiments, such as analyses of COVID-19 lockdowns [16], offer valuable real-world insights into the effects of emission reduction strategies. Nonetheless, challenges remain in integrating heterogeneous data sources, handling missing data, aligning spatial and temporal resolutions, and ensuring model transparency and generalizability[16], [17].

5. Conclusion

This review indicates the growing application of statistical and machine learning models in understanding and predicting the health effects of air pollution. Traditional models, such as GLM, GAM, and Poisson regression, remain valuable for clear interpretation, particularly in time-series and single-pollutant studies. However, modern methods such as Random Forest, XGBoost, LSTM, and hybrid models offer enhanced accuracy for complex, nonlinear, and large-scale data analysis. Bayesian models help address uncertainty and spatial variability. Combining satellite data with ground pollution measurements and hospital records enhances reliability. Preprocessing steps, such as lag selection and variable ranking, are crucial for achieving optimal model performance. Model choice should depend on the research aim, the type of data, and the resources available. This review demonstrates that no single model is suitable for all scenarios. Future efforts should focus on developing explainable hybrid models and enhancing data integration, ultimately creating region-specific, accurate, and adaptable models to inform public health decisions.

Compliance with ethical standards

Acknowledgments

The authors sincerely thank the Institute of Science, Nagpur, and its director for providing the necessary resources and support to conduct this research.

Disclosure of conflict of interest

The authors declare that there is no conflict of interest.

References

[1] Z. Kazemi et al., "Estimating the health impacts of exposure to Air pollutants and the evaluation of changes in their concentration0020using a linear model in Iran," Toxicol. Rep.., vol. 12, pp. 56–64, Jun. 2024, doi: 10.1016/j.toxrep.2023.12.006.

[2] L. Zhang et al., "Improving the construction and prediction strategy of the Air Quality Health Index (AQHI) using machine learning: A case study in Guangzhou, China," Ecotoxicol. Environ. Saf., vol. 287, p. 117287, Nov. 2024, doi: 10.1016/j.ecoenv.2024.117287.

[3] Y. Liu, L. Wen, Z. Lin, C. Xu, Y. Chen, and Y. Li, "Air quality historical correlation model based on time series," Sci. Rep.., vol. 14, no. 1, p. 22791, Oct. 2024, doi: 10.1038/s41598-024-74246-2.

[4] D.-Y. Jung et al., "Effect of Traffic-Related Air Pollution on Allergic Disease: Results of the Children's Health and Environmental Research," Allergy Asthma Immunol. Res., vol. 7, no. 4, p. 359, 2015, doi: 10.4168/aair.2015.7.4.359.

[5] D. Lee and G. Shaddick, "Modelling the effects of air pollution on health using Bayesian dynamic generalised linear models," Environmetrics, vol. 19, no. 8, pp. 785–804, Dec. 2008, doi: 10.1002/env.894.

[6] M. Pirani, N. Best, M. Blangiardo, S. Liverani, R. W. Atkinson, and G. W. Fuller, "Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles," Environ. Int., vol. 79, pp. 56–64, Jun. 2015, doi: 10.1016/j.envint.2015.02.010.

[7] G. Huang, D. Lee, and M. Scott, "An integrated Bayesian model for estimating the long-term health effects of air pollution by fusing modelled and measured pollution data: A case study of nitrogen dioxide concentrations in Scotland," Spat. Spatio-Temporal Epidemiol., vol. 14–15, pp. 63–74, Jul. 2015, doi: 10.1016/j.sste.2015.09.002.

[8] M. Blangiardo, M. Pirani, L. Kanapka, A. Hansell, and G. Fuller, "A hierarchical modelling approach to assess multi-pollutant effects in time-series studies," PLOS ONE, vol. 14, no. 3, p. e0212565, Mar. 2019, doi: 10.1371/journal.pone.0212565.

[9] B. S. Pimpale and A. A. Pandit, "Multioutput Ensemble Machine Learning Algorithm: A Prediction Model of Acute Respiratory Infection and Pneumonia Occurrence," Indian J. Sci. Technol., vol. 16, no. 45, pp. 4141–4155, Dec. 2023, doi: 10.17485/IJST/v16i45.1011.

[10] D. P. Singh, "An Extensive Analysis of Machine Learning Techniques for Predicting the Onset of Lung Cancer," Tuijin Jishu/Journal of Propulsion Technology, vol. 45, No.4 (2024).

[11] L. Liu et al., "Spatial–Temporal Analysis of Air Pollution, Climate Change, and Total Mortality in 120 Cities of China, 2012–2013," Front. Public Health, vol. 4, Jul. 2016, doi: 10.3389/fpubh. 2016.00143.

[12] Di. Zhong, "Contribution of Ambient Air Pollution on Risk Assessment of Type 2 Diabetes Mellitus via Explainable Machine Learning*," no. 2023.

[13] Sanjeev Prakashrao Kaulgud, "Leveraging Enhanced PSO and Proving Random Forest's Dominance for Prediction of Lung Cancer Severity," J. Inf. Syst. Eng. Manag., vol. 10, no. 3, pp. 60–70, Mar. 2025, doi: 10.52783/jisem.v10i3.3663.

[14] S. H. Chan et al., "Long-Term Air Pollution Exposure and Blood Pressure in the Sister Study," Environ. Health Perspect., vol. 123, no. 10, pp. 951–958, Oct. 2015, doi: 10.1289/ehp.1408125.

[15] T. G. Mason, C. M. Schooling, K. P. Chan, and L. Tian, "An evaluation of the air quality health index program on respiratory diseases in Hong Kong: An interrupted time series analysis," Atmos. Environ., vol. 211, pp. 151–158, Aug. 2019, doi: 10.1016/j.atmosenv.2019.05.013.

[16] A. Mohammadi, E. Pishgar, and J. Aguilera, "Spatial Prediction of High-Risk Areas for Asthma in Metropolitan Areas: A Machine Learning Approach Applied to Tehran, Iran," ISPRS Int. J. Geo-Inf., vol. 14, no. 3, p. 105, Mar. 2025, doi: 10.3390/ijgi14030105.

[17] A. Chauhan, G. P. Sai, and C.-Y. Hsu, "Advanced statistical analysis of air quality and its health impacts in India: Quantifying significance by detangling weather-driven effects," Heliyon, vol. 11, no. 2, p. e41762, Jan. 2025, doi: 10.1016/j.heliyon.2025.e41762.

[18] M. Schwarz et al., "Temporal variations in the short-term effects of ambient air pollution on cardiovascular and respiratory mortality: a pooled analysis of 380 urban areas over 22 years," Lancet Planet. Health, vol. 8, no. 9, pp. e657-e665, Sep. 2024, doi: 10.1016/S2542-5196(24)00168-2.

[19] Ayesha et al., "Modelling health outcomes of air pollution in the Middle East by using support vector machines and neural networks," Sci. Rep.., vol. 14, no. 1, p. 21517, Sep. 2024, doi: 10.1038/s41598-024-71694-8.

[20] A. Amer, N. Mushtaq, O. Albalawi, M. Hanif, E. E. Mahmoud, and M. Nabi, "Forecasting mortality and DALYs from air pollution in SAARC nations," Sci. Rep.., vol. 14, no. 1, p. 25898, Oct. 2024, doi: 10.1038/s41598-024-76760-9.

[21] S. Ram, W. Zhang, M. Williams, and Y. Pengetnze, "Predicting Asthma-Related Emergency Department Visits Using Big Data," IEEE J. Biomed. Health Inform., vol. 19, no. 4, pp. 1216–1223, Jul. 2015, doi: 10.1109/JBHI.2015.2404829.

[22] W. Jin, "The Relationship between Lung Cancer Prevalence and Air Quality and Other Factors," Highlights Sci. Eng. Technol., vol. 99, pp. 34–41, Jun. 2024, doi: 10.54097/3htthh72.

[23] A. Kim, J. Jung, J. Hong, and S.-J. Yoon, "Time series analysis of meteorological factors and air pollutants and their association with hospital admissions for acute myocardial infarction in Korea," Int. J. Cardiol., vol. 322, pp. 220–226, Jan. 2021, doi: 10.1016/j.ijcard.2020.08.060.