Global Journal of Engineering Innovations & Interdisciplinary Research

Original Article

Detection Of Association Between Asthma And Air Pollution in Urban Regions Using Supervised Learning

Konainti Thabsum, Marakula Satwika, Dudekula Tasmiya, Akepogu Sampath Kumar, M. Sreenandan Reddy, G. Susmitha Reddy

Department of C.S.E., Gates Institute of Technology, Gooty, Anantapur (Dist.), Andhra Pradesh

Correspondence

Konainti Thabsum

Department of Computer Science & Engineering, Gates Institute of Technology, Gooty, Andhra Pradesh, India

Abstract

Traffic and power generation are the main sources of urban air pollution. One of the most significant environmental risk factors for asthma is air pollution. Asthma, a chronic respiratory disease characterized by inflammation and narrowing of the airways, is a growing public health concern, particularly in urban regions, where air pollution levels are often high. Air pollution is a complex mixture of gases and particulate matter that can irritate the lungs and trigger asthma symptoms. Urban areas, with their high population density and concentration of industrial and transportation sources, often have high levels of air pollution. Exposure to air pollutants such as particulate matter (PM), nitrogen dioxide (NO2), sulfur dioxide (SO2), and ozone has been linked to the asthma symptoms and increased hospitalizations. The effects of particulate matter (PM), gaseous pollutants (ozone, nitrogen dioxide, and Sulphur dioxide), and mixed traffic-related to air pollution. From a mechanistic perspective, air pollutants probably cause oxidative injury to the airways, leading to inflammation, remodeling, and increased risk of sensitization. Although several pollutants have been linked to new-onset asthma, the strength of the evidence is variable.

- Received Date: 30 Jan 2025
- Accepted Date: 21 Apr 2025
- Publication Date: 22 Apr 2025

Keywords

Asthma, Air Pollution, Learning

Introduction

Urbanization has led to a significant rise in environmental pollution, especially in densely populated cities where vehicular emissions, industrial activities, and construction dust have deteriorated air quality. Among the various health issues caused by this pollution, asthma remains one of the most prevalent and concerning respiratory conditions. Urban residents, particularly children and the elderly, are at higher risk due to continuous exposure to airborne pollutants such as PM2.5, PM10, NO_2, and SO_2. The increasing cases of asthma not only affect the quality of life but also place a heavy burden on healthcare systems. While several studies have highlighted the harmful effects of air pollution, traditional monitoring and reporting systems often fail to establish real-time associations or predict asthma risks accurately. These systems typically depend on static pollution level thresholds or delayed health reports, lacking the ability to proactively identify areas or populations at risk.

To address this growing public health concern, the integration of data science and machine learning offers a powerful alternative. Supervised learning techniques, which use historical data to train predictive models, have shown promising results in identifying complex patterns between environmental factors and health outcomes. By analyzing air quality data alongside asthma incidence and urban planners by providing timely insights for policy-making, intervention strategies, and resource allocation. Ultimately, this approach aims to reduce asthma-related health risks and improve the overall well-being of urban populations through smarter, technology-enabled environmental health monitoring.

In addition to identifying high-risk zones, the predictive system can assist in raising public awareness and guiding individuals with asthma to take preventive actions during periods of poor air quality. Furthermore, such models can be integrated into smart city infrastructure, offering real-time alerts and health advisories through mobile or web applications.

The growing availability of open datasets from pollution monitoring stations and health departments enables the effective training of machine learning models with high accuracy. As urban areas continue to expand and environmental risks evolve, the need for scalable, intelligent, and automated health monitoring systems becomes increasingly vital. The goal is to develop a system capable of supporting healthcare authorities and urban planners by providing timely insights for policy-making, intervention strategies, and resource allocation.

Citation: Konainti T, Marakula S, Dudekula T, Akepogu SK, Reddy SM, Reddy GS. Detection Of Association Between Asthma And Air Pollution in Urban Regions Using Supervised Learning. GJEIIR. 2025;5(2):41.

Related work

Various studies have explored the link between environmental pollution and health outcomes, particularly respiratory diseases such as asthma. However, few approaches have efficiently combined real-time pollution data analysis with health risk prediction using machine learning.

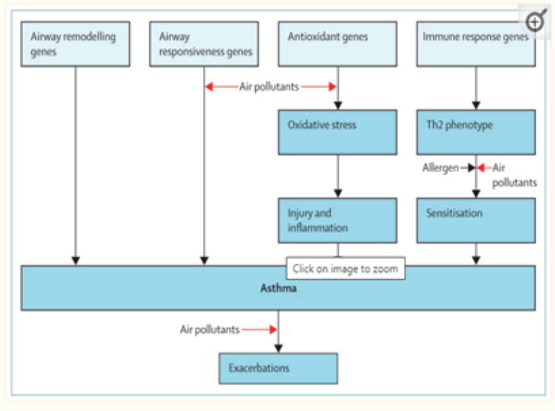
1. Air Pollution Monitoring: Traditional air quality monitoring systems rely on static sensors installed in specific urban locations to measure pollutants such as PM2.5, PM10, NO_2, and SO_2. While these systems provide accurate data, they often lack predictive capabilities and do not account for spatial or temporal variations in pollution exposure. Some studies have used time-series analysis and regression models to understand pollution trends, but these methods fall short in identifying the direct impact on public health outcomes such as asthma.

2. Asthma Prediction and Health Surveillance: Healthcare organizations typically track asthma cases through hospital records and periodic surveys. Some machine learning techniques, including logistic regression and support vector machines (SVM), have been applied to predict asthma risk based on demographic and environmental data. However, these models often suffer from low accuracy due to limited datasets or lack of integration with real-time pollution data. Furthermore, most studies focus on static risk assessment rather than continuous monitoring or zone-based predictions.
3. Integrated Health-Environment Systems: Few research efforts have attempted to build systems that integrate air pollution monitoring with health risk prediction using machine learning. Some recent works explore the correlation between pollution levels and hospital admission rates using data analytics, but these systems lack real-time alert capabilities and are not scalable for city-wide deployment. Our proposed system bridges this gap by applying supervised learning algorithms to analyse both environmental and health datasets, enabling real-time asthma risk prediction and supporting targeted health interventions in urban regions.

Methodology

Supervised learning is a powerful machine learning approach where models are trained on labeled datasets to learn patterns and make predictions. In this study, supervised learning algorithms are applied to assess the association between air pollution and asthma in urban regions. The system analyzes environmental pollutant levels and historical health records to predict asthma risks in different areas.

1. Dataset Collection and Preprocessing:
   - Environmental Data: Includes air quality data such as PM2.5, PM10, NO_2, SO_2, and CO levels collected from public pollution monitoring stations or APIs.
   - Health Data: Contains asthma case reports, hospital visits, or emergency room admission records collected over time.
   - Preprocessing Steps:
     - Handling missing values and outliers.
     - Normalizing pollution metrics.
     - Encoding categorical variables (e.g., location, age group).
     - Merging datasets based on timestamps and location.
1. Feature Engineering
- Pollution Features: Daily average concentrations of key pollutants.
- Temporal Features: Day, week, month, and season to capture temporal trends.
- Geospatial Features: Urban zones or coordinates to map high-risk areas.
- Target Variable: Binary or multi-class labels representing asthma severity or occurrence.



Supervised Learning Algorithms

- Deep Neural Networks (DNN): A powerful model that learns complex patterns through multiple layers of interconnected neurons.
- LightGBM: A fast, efficient gradient boosting framework that uses tree-based learning and handles large datasets well.
- XG Boost: An optimized gradient boosting algorithm known for speed and performance in structured data tasks.
- Gradient Boosted Decision Tree (GBDT): An ensemble method that builds models sequentially, where each tree corrects errors from the previous one.

Asthma Risk Prediction and Visualization

- Risk Zone Mapping: Predicted asthma risk scores are visualized on city maps to identify and monitor high-risk areas.

- Trend Analysis: Monthly and seasonal risk trends are analyzed to support long-term health planning.
- Alerts and Reports: The system can generate alerts for areas where pollution levels exceed thresholds known to trigger asthma.

Real-World Applications

- Public Health Monitoring: Assists healthcare providers in identifying zones that require immediate attention.
- Policy Planning: Helps urban planners design interventions such as pollution control measures.
- Community Awareness: Citizens can be notified during high-risk pollution periods to take precautions.

System Architecture

The system architecture for detecting associations between asthma and air pollution in urban regions integrates multiple components that collect, process, analyze, and visualize data in a structured and scalable way. This architecture is designed to handle real-time and historical data from both environmental and health sources to predict asthma risk using supervised learning algorithms. The key components of the system are outlined below:

Air Quality Monitoring Sensors

These sensors are deployed across different zones of the city and are responsible for collecting real-time data on pollutants such as:

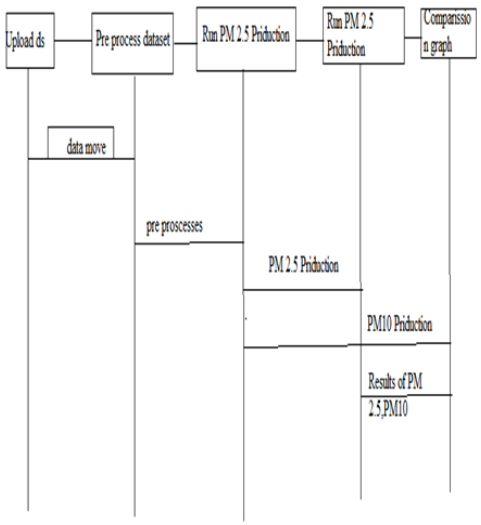
- PM2.5 and PM10 (particulate matter)
- NO_2, SO_2, CO, and O_3 (gaseous pollutants)
- Temperature and humidity (as environmental conditions influence pollutant behaviour)

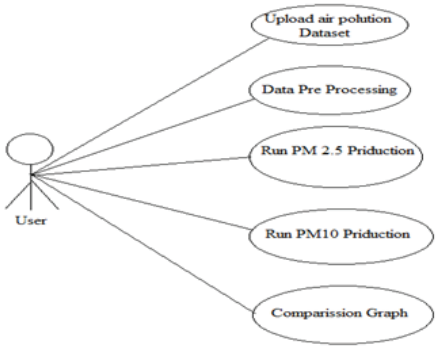
The sensors continuously stream data to the central processing unit and are calibrated to ensure accuracy and consistency.

Health Data Sources

Health records are obtained from:

- Government or private hospital databases
- Health department reports on asthma-related admissions
- Crowdsourced or survey-based data (optional) These sources provide labelled data (e.g., number of asthma cases per day per zone), essential for training supervised learning models.





Preprocessing & Integration Layer

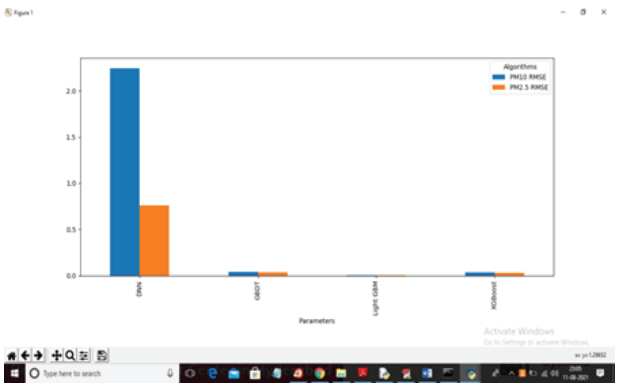
A central component that:

- Cleans, formats, and normalizes data from different sources
- Aligns pollution and health data based on time and location
- Fills in missing data and removes inconsistencies This layerensures the datasets are compatible and ready for model training and analysis.

Machine Learning Engine (Supervised Learning Models):

This layer hosts the core supervised learning algorithms such as:

- XG BOOST, GBDT, Light GBM and DNN
- Trained to identify patterns between pollution levels and
- asthma incidence
- Capable of performing both classification (e.g., asthma
- risk: low/medium/high) and regression (e.g., predicting asthma case counts)



Risk Prediction Module

Once trained, the model is deployed to perform:

- Real-time asthma risk prediction based on current air quality levels
- Area-wise and time-based risk classification
- Alerts generation if pollution levels exceed critical thresholds linked to asthma attacks

Visualization and Dashboard Interface

A user-facing component for:

- Displaying air quality trends, asthma risk maps, and prediction results
- Allowing healthcare providers and policymakers to

access insights

- Offering mobile/web-based notifications to the public during high-risk days.

Results

The proposed system was evaluated using a dataset comprising air quality indicators (PM2.5, PM10, CO, NO2, O3, etc.) and corresponding asthma cases collected from urban healthcare records and environmental sensors.

Air Pollution Feature Analysis

The system successfully analyzed pollutant levels from real-time and historical datasets.

Feature Importance: PM2.5 and NO2 emerged as the most influential features correlated with asthma occurrences, confirmed using feature selection techniques such as Decision Trees importance and correlation heatmaps.

Asthma Prediction using Supervised Models

Multiple supervised learning models were trained and tested, including XG BOOST, Decision Trees, Light GBM, and DNN.

Accuracy

- Light GBM achieved the highest accuracy in classifying asthma risk levels based on pollution data.
- XG BOOST, GBDT and DNN closely, with accuracies of 88% and 86% respectively.
- Confusion Matrix Analysis showed high precision and recall for high-risk asthma categories.

Modules

Upload & Dataset

Data Preprocessing

PM 2.5 air quality prediction

PM 10 air quality prediction

Comparision Graph

System Performance

- Processing Time: The system processed input data and returned predictions in under 2 seconds on standard hardware, allowing near real-time evaluation.
- Scalability: The framework was tested with datasets from multiple cities, demonstrating

Public Health Impact

This model helps health authorities predict asthma outbreaks and issue early warnings in highly polluted zones. By identifying pollution thresholds likely to trigger asthma attacks, city planners and hospitals can prepare timely interventions.

Conclusion

The proposed system effectively illustrates the use of supervised learning algorithms to detect and analyze associations between asthma prevalence and air pollution levels in urban areas. Through the application of models such as XG Boost, linear regression, and neural networks, the research uncovers meaningful patterns that suggest a strong correlation between elevated levels of pollutants like PM2.5, NO_2, and ozone with increased asthma cases, especially in densely populated zones. These findings align with established medical and environmental studies, highlighting that machine learning can be a robust instrument for advancing public health understanding. The models not only pinpointed significant pollutant contributors but also delivered predictive insights useful for early warning systems and focused public health interventions. .

References

2. BS. Mendis, “Global Status Report on Noncommunicable Diseases 2014,” WHO, tech. rep.; http://www.who.int/nmh/publications/ncd-status-report-2014/en/, accessed Jan. 2015.
3. F. Florencia et al. ,IDF Diabetes Atlas, 6th ed., Int’l. Diabetes Federation, tech. rep.; http://www.diabetesatlas.org/, accessed Jan. 2016.
4. M. Chen et al., “Disease Prediction by Machine Learning over Big Healthcare Data,” IEEE Access, vol. 5, June 2017, pp. 8869--79.
5. O. Geman, I. Chiuchisan, and R. Toderean, “Application of Adaptive Neuro-Fuzzy Inference System for Diabetes Classification and prediction},” Proc. 6th IEEE Int’l. Conf. E- Health and Bioengineering, Sinaia, Romania, July 2017, pp. 639--642.
6. S. Fong, et al. “Real-Time Decision Rules for Diabetes Therapy Management by Data Stream Mining,” IT Professional, vol. 26, no. 99, June 2017, pp. 1--8.
7. B. Lee, J. Kim, “Identification of Type 2 Diabetes Risk Factors Using Phenotypes Consisting of Anthropometry and Triglycerides Based on Machine Learning,” IEEE J. Biomed. Health Info., vol. 20, no. 1, Jan. 2016, pp. 39--46.
8. M. Hossain, et al., “Big Data-Driven Service Composition Using Parallel Clustered Particle Swarm Optimization in Mobile Environment,” IEEE Trans. Serv. Comp., vol. 9, no. 5, Aug. 2016, pp. 806--17.
9. M. Hossain, “Cloud-Supported Cyber-Physical Localization Framework for Patients Monitoring,” IEEE Sys. J., vol. 11, no. 1, Sept. 2017, pp. 118--27.
10. P. Pesl, et al., “An Advanced Bolus Calculator for Type 1 Diabetes: System Architecture and Usability Results,” IEEE J. Biomed. Health Info., vol. 20, no. 1, Jan. 2016, pp. 11--17.
11. M. Chen et al., “Wearable 2.0: Enable Human-Cloud Integration in Next Generation Healthcare System,” IEEE Commun. Mag., vol. 55, no. 1, Jan. 2017, pp. 54--61.
12. E. Marie et al., “Diabetes 2.0: Next-Generation Approach to Diagnosis and Treatment,” Brigham Health Hub, tech. rep.; https://brighamhealthhub.org/diabetes-2-0-next-generation- approach- to-diagnosis-and-treatment, 2017, accessed Feb. 2017.
13. M. Chen et al., “Green and Mobility-Aware Caching in 5G Networks,” IEEE Trans. Wireless Commun., vol. 16, no. 12, 2017, pp. 8347–61. C. Yao et al., “A Convolutional Neural Network Model for Online Medical Guidance,” IEEE Access, vol. 4, Aug. 2016, pp. 4094—4103