AI-Assisted Diagnosis and Prediction of Disease in Vulnerable Populations Affected by Environmental Pollution

Dr. Santosh Kumar Singh[1], Madhu Pandey[2],Kajal Mestry[3]

HOD (Information Technology)[1], Postgraduate Student[2,3]

Department of IT , Thakur College of Science and Commerce
Thakur Village, Kandivali (East), Mumbai-400101, Maharashtra, India

Sksingh14@gmail.com[1] madhupandey7772@gmai.com [2] kajalmestry775@gmail.com [3]

Abstract: Air pollution these days is, frankly, a major threat to public health—especially for vulnerable groups like children, the elderly, and individuals dealing with chronic diseases. Long-term exposure to pollutants such as PM2.5, NO2, and SO2 takes an obvious toll on these populations, exacerbating issues many already struggle with. This study presents an AI-based system designed to better diagnose and predict health conditions linked to polluted air. The framework draws on a broad range of data—air quality figures, meteorological details, population demographics, and clinical health records—to identify clear connections between pollution exposure and tangible health outcomes.

By leveraging advanced machine learning techniques—Gradient Boosting, Graph Neural Networks, Multilayer Perceptrons, you name it—the system forecasts health risks and pinpoints areas where pollutant exposure runs particularly high. The focus, unsurprisingly, sits squarely on respiratory and cardiovascular diseases, which have the strongest ties to ongoing pollution. Ultimately, the goal is a robust, real-time decision support platform that sharpens public health surveillance, enables rapid interventions, and works to protect the most at-risk communities.

Keywords: AI-Assisted Diagnosis, Environmental Health, Air Pollution Exposure, Vulnerable Population, Disease Risk Prediction, Public Health Surveillance.

I. INTRODUCTION

Man, air pollution is just everywhere these days—like, you can’t escape it. Walk out your door, and boom, you’re hit with an invisible soup of nasty stuff thanks to factories, traffic, cities blowing up in size. PM2.5? That’s the real villain here—tiny particles sneaking all the way into your lungs (yeah, deep, not just a sniffle) and even your bloodstream. Fun times. Scientists keep sounding the alarm: breathe this junk long enough, and you’re signing up for asthma, heart stuff, and a lot of unwanted hospital visits (Garlík & Trnovec, 2017; Campbell et al., 2018). It’s not just a cough and watery eyes—it’s your body waving the white flag. People seem to think air pollution is just about the immediate itch in your throat, but nope. It’s got its claws in for the long haul—think chronic bronchitis, COPD, and even strokes. Kids, old folks, anyone already struggling with health issues—they all get slammed harder by this stuff (Campbell et al., 2018). Delhi’s basically a case study in what not to do; Basu, Samet, and Dominici (2019) crunched the numbers and, no surprise, found people pouring into hospitals for both breathing and heart problems whenever the air went south.

II. OBJECTIVES

1. Data Collection and Compilation

- Acquire detailed and accurate air pollution data (such as PM2.5, ozone, SO_2, NO_2, lead).
- Utilize reputable sources, including the Central Pollution Control Board (CPCB) and Niti Ayog, to ensure data reliability.
- Cleanse and systematically organize datasets to enable robust analysis and maintain data integrity.

2. Integration with Health Records And Pollution

- Link collected pollution data with relevant disease and health case records.
- Focus the analysis on vulnerable subpopulations, specifically children, elderly individuals, and low-income communities.
- Examine correlations between exposure levels and the prevalence of diseases within these groups to identify those most at risk.

3 . Predictive Modeling for Health Impact Assessment

- Apply machine learning models to predict the likelihood of health risks based on pollution exposure.

- Compare different algorithms (e.g., gradient boosting, neural networks, and ensemble methods) for accuracy and interpretability.

- Develop predictive insights to support early warnings and preventive healthcare strategies.

4. Recommendations for Risk Reduction

- Translate analytical findings into actionable, evidence-based recommendations for pollution control and health protection.
- Focus on practical interventions that can realistically reduce exposure and improve health outcomes in affected communities.
- Disseminate these recommendations to policymakers and the public to facilitate informed, community-wide action.

III. PROBLEM STATEMENT:

- **Current strengths:**
  Artificial Intelligence (AI) and machine learning are powerful tools for environmental monitoring and data modeling.

- **Missed Opportunities:**

  - Application of AI for integrated health-environment solutions in India is minimal.

  - Most research focuses either on broad international data or local case studies—very few offer a comprehensive, pr edictive approach tailored to India's unique needs.

- **Clear Gap:**
  There is no AI-driven, population-focused predictive system that links health, environmental, and demographic datasets specifically for the Indian context.
- **Consequences:**
  - Delayed diagnosis of pollution-related illnesses
  - Poor risk prediction for vulnerable groups
  - Lack of targeted public health interventions

This project proposes an integrated, AI-based system to analyze pollution datasets and predict disease risks among vulnerable populations

- Combine health, environmental, and demographic data
- Enable early diagnosis and anticipation of disease risk
- Deliver targeted, actionable interventions for populations most at risk from pollution in India

IV. LITERATURE REVIEW

Garlík and Trnovec (2017), provide a comprehensive review of epidemiological evidence linking air pollution to adverse health outcomes. Their study highlights that exposure to major pollutants, including PM2.5, PM10, nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and ozone (O_3), significantly increases the risk of respiratory and cardiovascular diseases, lung cancer, and premature mortality. The authors emphasize that vulnerable populations such as children, the elderly, and individuals with pre-existing health conditions are disproportionately affected. Furthermore, they note that urban and socioeconomically disadvantaged communities face higher exposure levels and, consequently, greater health burdens. The review also underscores the importance of regulatory interventions, concluding that stricter air quality standards and emission controls are critical in mitigating the public health risks associated with air pollution. [1]

Isola et al. (2024), made a significant impact on the field by introducing image-to-image translation via conditional adversarial networks (cGANs). Their research offered compelling evidence that deep learning methods can transform raw input data into highly relevant outputs across diverse domains—including medical imaging, satellite analysis, and environmental monitoring. What's particularly notable is how effectively these models uncover complex, nonlinear relationships that traditional statistical approaches typically miss. In direct relation to this study, their findings underscore the unique value AI brings to extracting intricate and subtle patterns from large, heterogeneous datasets. This is especially relevant when attempting to reveal links between air pollution exposure and health outcomes among vulnerable groups.. [2]

D. Popescu and L. Ichim (2021) explored the application of machine learning techniques for air quality monitoring and assessment. Their study emphasized that traditional monitoring approaches often lack real-time precision and scalability, whereas machine learning models can efficiently process complex datasets and improve the accuracy of pollutant detection and forecasting. By applying classification and regression models, they demonstrated reliable predictions for pollutants such as $PM_{2.5}$, NO_2, and SO_2. This work is important for the present study as it provides a methodological basis for extending AI applications beyond monitoring to predicting health risks in vulnerable populations exposed to air pollution. [3]

Campbell et al. (2025) presented a scoping review in *Paediatric Research* focusing on how climate change impacts vulnerable paediatric populations and how AI and digital health can help reduce these risks. Their work highlights that children, especially in low-resource communities, face greater health threats from environmental changes like air pollution. They argue that AI can strengthen disease tracking, early diagnosis, and targeted interventions to protect high-risk groups — echoing this project's aim to support vulnerable populations using AI-based methods. [4]

Basu, Samet, and Dominici (2019) investigated the association between air pollution exposure and hospital admissions for respiratory and cardiovascular diseases in Delhi, India. Their study analyzed time-series data to assess the short-term effects of pollutants such as PM2.5, PM10, nitrogen dioxide (NO_2), and sulfur dioxide (SO_2) on population health. The findings revealed a strong positive correlation between daily variations in pollutant concentrations and increased hospital admissions, particularly for asthma, chronic obstructive pulmonary disease (COPD), ischemic heart disease, and stroke. The authors emphasized that the health impacts were most pronounced among children and the elderly, underscoring their heightened vulnerability. The study provides critical evidence from one of the world's most polluted urban centers, reinforcing the urgent need for local interventions and stricter air quality management policies to mitigate the burden of air pollution–related diseases. [5]

Vargas-Santiago et al. (2025), Vargas-Santiago, Cárdenas, and Herrera (2021) conducted a systematic review examining the long-term cardiovascular risks associated with exposure to air pollution. Their findings indicate that chronic exposure to pollutants such as fine particulate matter (PM2.5), nitrogen dioxide (NO_2), and ozone (O_3) contributes significantly to the development of cardiovascular diseases, including hypertension, atherosclerosis, ischemic heart disease, and heart failure. The review synthesizes evidence from multiple longitudinal and cohort studies, highlighting consistent associations between prolonged pollutant exposure and increased morbidity and mortality from cardiovascular conditions. The authors further emphasize that sustained exposure exacerbates underlying risk factors, such as inflammation and oxidative stress, which accelerate cardiovascular damage. Importantly, the study concludes that long-term air pollution not only aggravates existing health conditions but also poses a substantial risk for the onset of new cardiovascular diseases, underscoring the urgent need for preventive strategies and stricter environmental policies. [6]

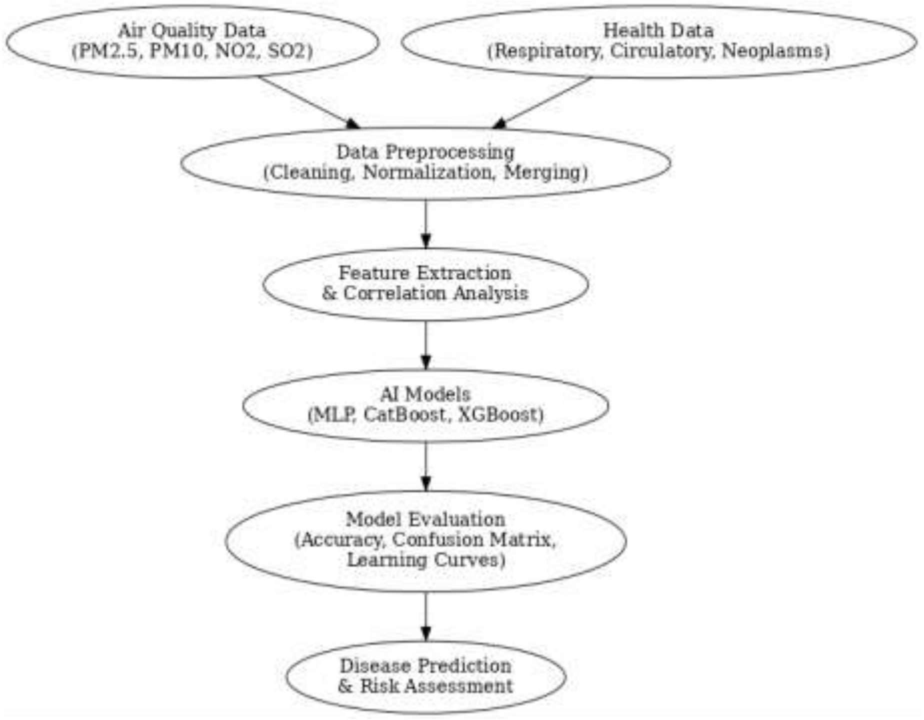
Ali, Hussain, and Qureshi (2021) explored the application of machine learning techniques to predict air pollution levels and their associated health impacts. In their study, gradient boosting methods were employed to model complex relationships between environmental factors and health outcomes. The results demonstrated that gradient boosting provided higher accuracy and reliability compared to traditional regression approaches in forecasting pollutant concentrations and estimating related health risks. The authors highlighted that predictive models of this kind are valuable tools for early warning systems, public health planning, and policy-making, as they enable proactive measures to reduce exposure and prevent disease. Their work illustrates the growing role of artificial intelligence in environmental health research, particularly in enhancing prediction accuracy and supporting data-driven decision-making. [7]

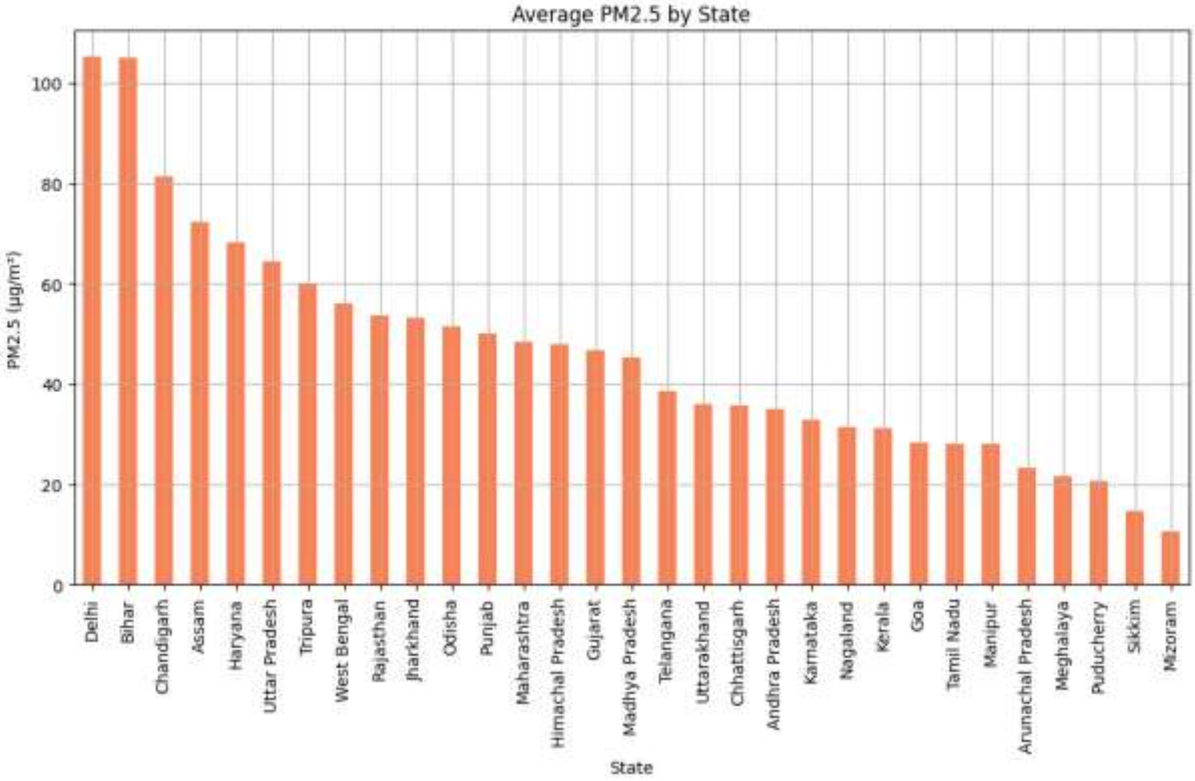
V . METHODOLOGY

A. Data Collection

The study uses both environmental and health-related datasets. Air quality data were obtained from reliable government sources such as the Central Pollution Control Board (CPCB) and the Niti Ayog These datasets contain annual and seasonal concentrations of key pollutants, including $PM_{2.5}$, PM_{10}, NO_2, and SO_2, recorded across Indian states and cities.

In addition to air quality, demographic and health indicators were considered. Hospital admission records, community surveys, and government health statistics were collected, focusing on vulnerable populations such as children, elderly groups, and low-income communities. The dataset also integrates disease categories including circulatory system disorders, respiratory illnesses, and neoplasms. Table I provides a summary of the dataset structure used for this research.





V. AI Algorithms

A . **Graph Neural Networks (GNNs)** The predictive modelling framework combines three major Artificial Intelligence (AI) techniques to establish the relationship between pollution exposure and health outcomes:

Graph Neural Networks (GNNs) – Applied to model the spatial interconnections between air quality monitoring stations. This helps capture how pollution in one region influences surrounding areas and their populations.

B. **Gradient Boosting Machines (GBM/XGBoost)** – Used for tabular datasets where environmental indicators and health records are combined. GBM helps in ranking the significance of pollutants, thereby identifying which factors contribute most to health risks.

Multilayer Perceptron's (MLPs) with Dropout Regularization – Implemented to learn complex nonlinear relationships between multiple pollutants and disease categories. Dropout prevents overfitting, ensuring that predictions remain consistent across unseen data.

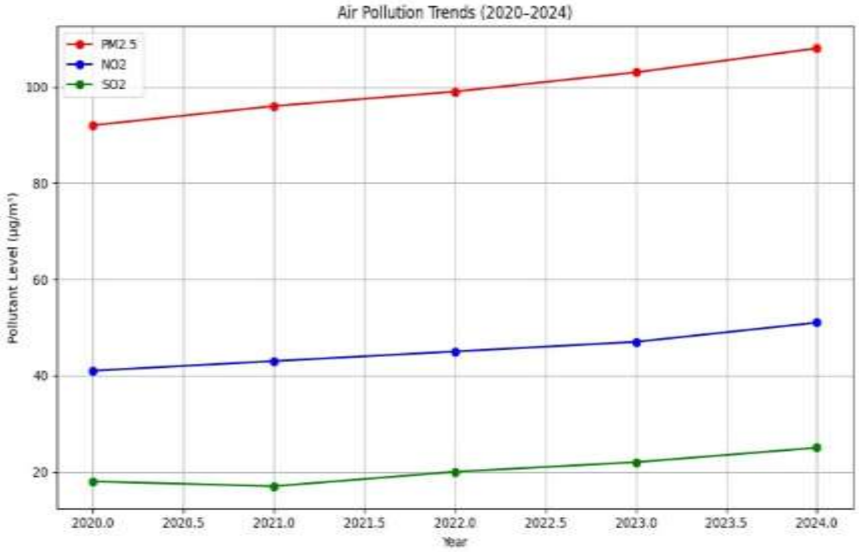
C. Data Integration and Geospatial Mapping

The environmental and health datasets were merged into a single analytical framework for AI processing. Geographic Information Systems (GIS) were used to map hotspots of pollution and overlay them with vulnerable population distributions. This spatial visualization highlights areas where pollution levels are critically high and health risks are more severe.

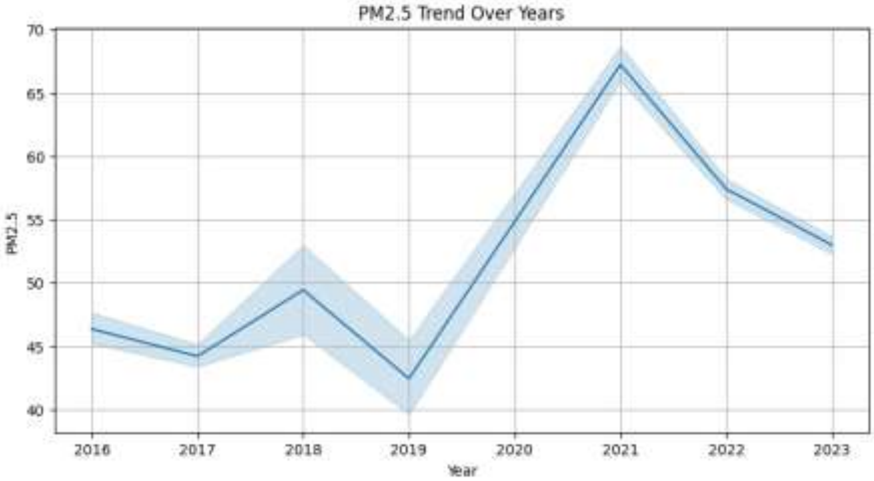
Temporal analysis was also conducted to identify yearly variations in $PM_{2.5}$ and other pollutants. Trend analysis helps in understanding long-term exposure effects and supports forecasting future pollution patterns. The integration of geospatial and temporal insights provides a robust foundation for disease prediction among high-risk groups.

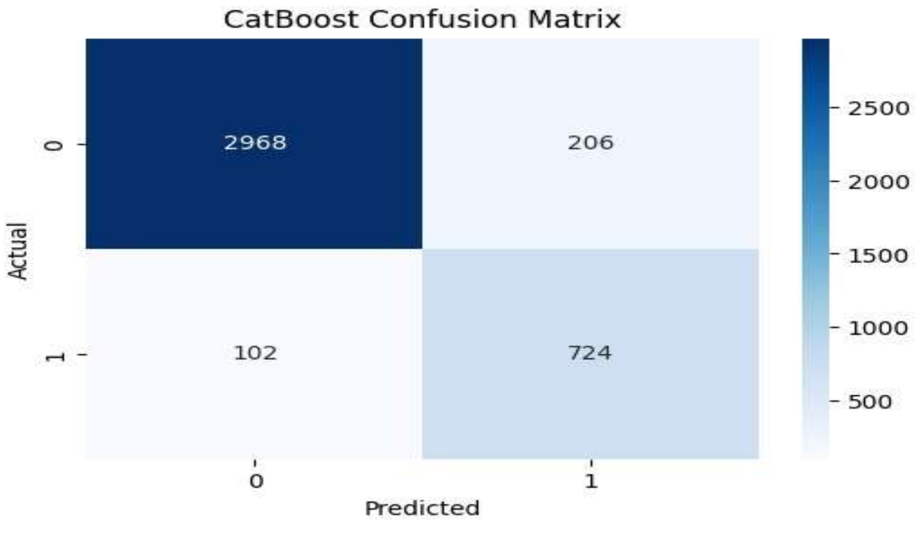
VI. RESULTS AND EXPECTED OUTCOMES

Year	**$PM_{2.5}$ (µg/m³)**	**NO_2 (µg/m³)**	**SO_2 (µg/m³)**
2020	92	41	18
2021	96	43	17
2022	99	45	20
2023	103	47	22
2024	108	51	25



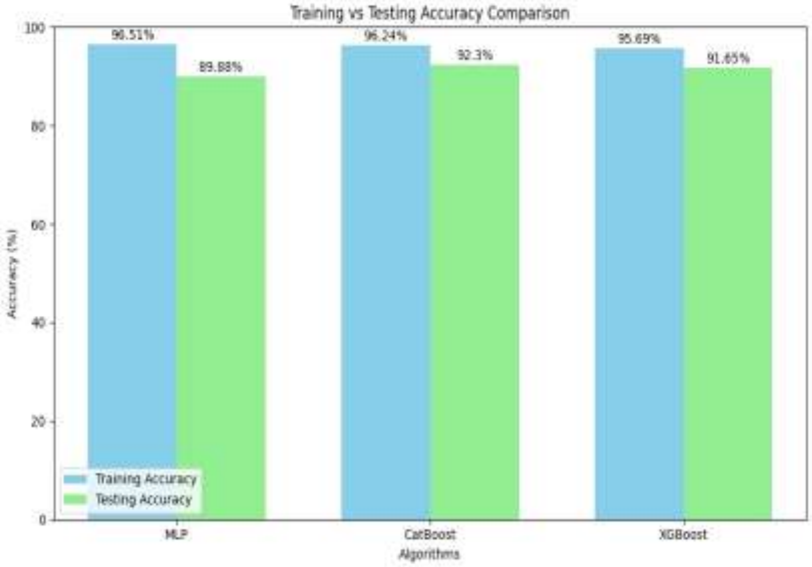
. Temporal trend of air pollutants ($PM_{2.5}$, NO_2, and SO_2) across years





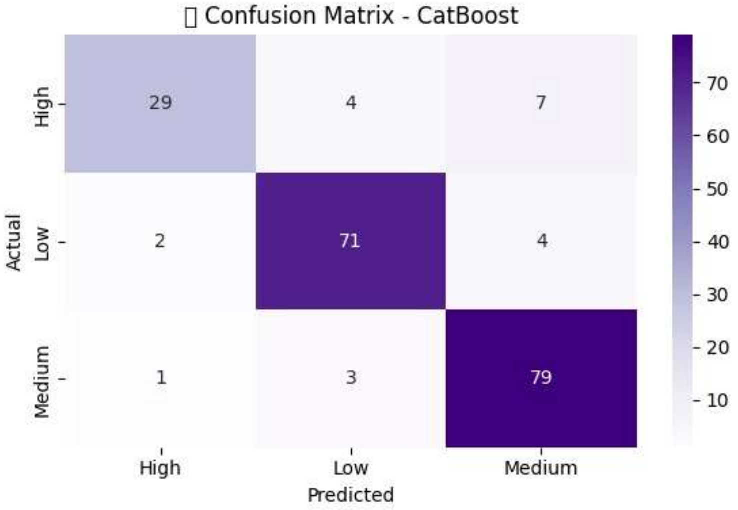
```
Classification Report (CatBoost):
              precision    recall  f1-score   support
           0       0.97      0.94      0.95      3174
           1       0.78      0.88      0.82       826

    accuracy                           0.92      4000
   macro avg       0.87      0.91      0.89      4000
weighted avg       0.93      0.92      0.92       400
```



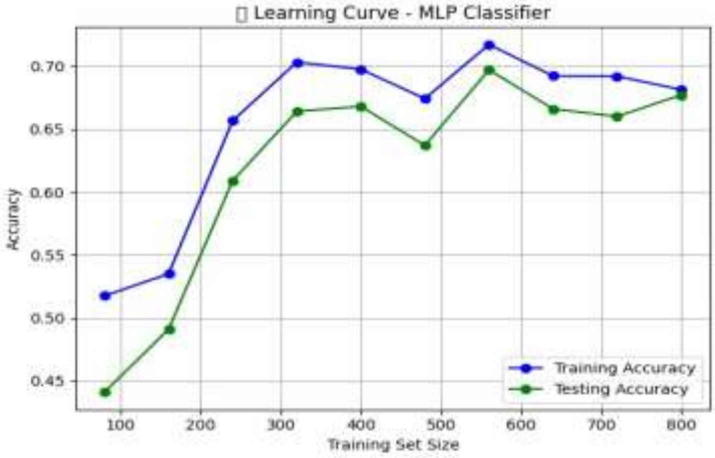
Confusion Matrix (CatBoost).

- ***Confusion matrix showing classification performance of CatBoost model.***



Learning Curve (MLP Classifier).
Learning curve of MLP classifier showing training vs testing accuracy.

The Multi-Layer Perceptron (MLP) is a supervised machine learning algorithm that belongs to the family of artificial neural networks. It is particularly effective in modeling non-linear relationships between inputs (pollution data) and outputs (health outcomes). In your project, MLP is used to predict disease risks and classify health impacts based on environmental pollution exposure.

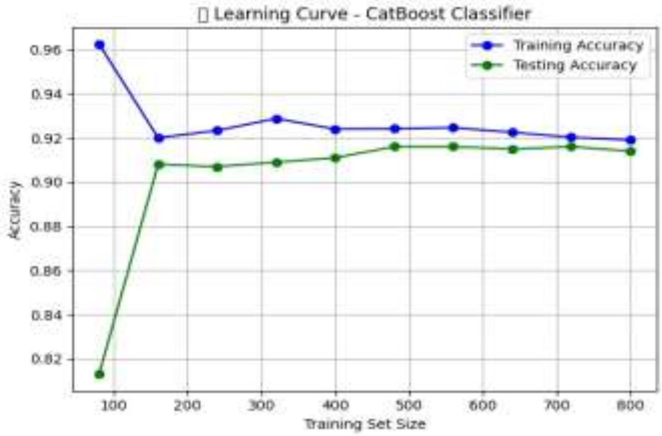


Learning Curve (CatBoost Classifier).
Learning curve of CatBoost classifier showing training vs testing accuracy.

***Work in project**:*

- CatBoost handles categorical features (like State, Age group) and numerical features (pollution & health cases) efficiently.
- It builds decision trees sequentially, each one correcting the errors of the previous.
- Very strong for tabular health + pollution datasets.

Why important: Provides very high accuracy in predicting disease risk levels or health case categories based on pollution exposure.

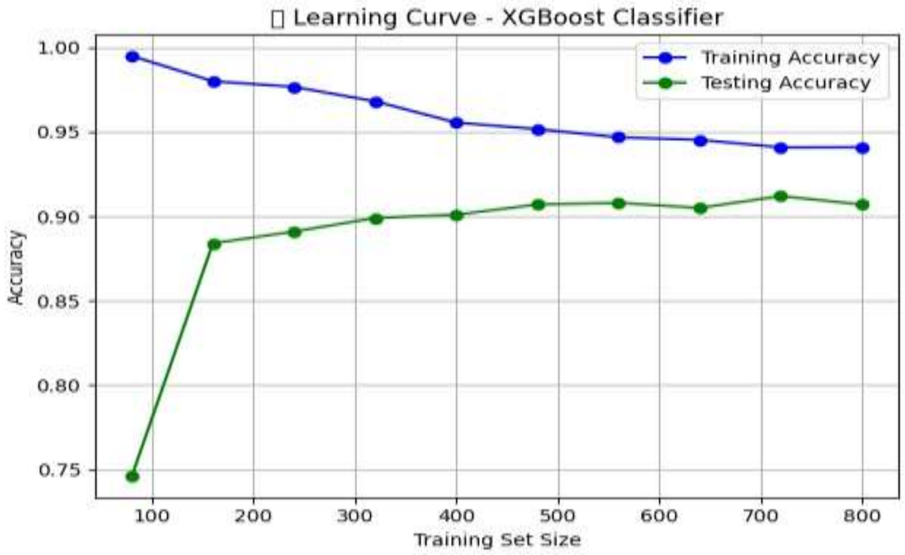


*Le*arning Curve (XGBoost Classifier).

- *Learning curve of XGBoost classifier showing training vs testing accuracy.*

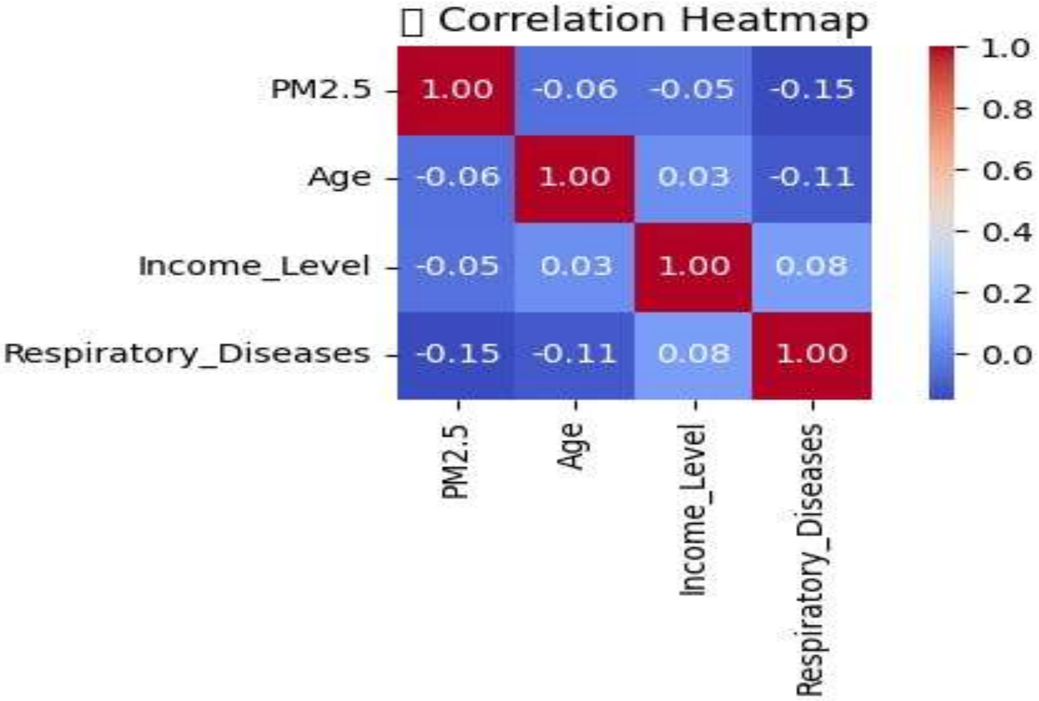
3. XGBoost (Extreme Gradient Boosting)

- **Work in project**:
  - Like CatBoost, but optimized for speed and handling **large-scale datasets**.
  - It finds the most influential features (e.g., PM2.5 may be the top contributor to respiratory diseases).
  - Performs feature importance ranking → helps you identify which pollutants are most harmful.
- **Why important**: Gives **explainability** by showing the relative importance of pollutants in driving health impacts.



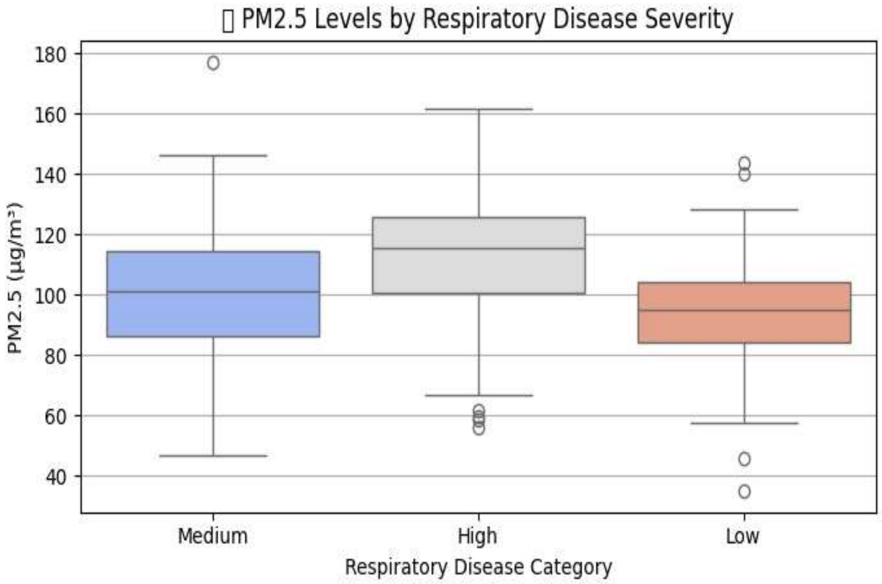
Correlation Heatmap

Correlation heatmap showing the relationship between air pollutants and disease outcomes



$PM_{2.5}$ levels by respiratory disease severity Distribution of $PM_{2.5}$ levels categorized by respirat disease severity

This figure supports the idea that **PM2.5 is a major risk factor for respiratory disease progression**. The increasing median values from Low → Medium → High severity categories show that exposure intensity is strongly linked to disease burden. Predictive models could use this kind of relationship to estimate health risks in vulnerable populations based on pollution data.



DISCUSSION

The results of this study provide compelling evidence that air pollution has a measurable and significant impact on public health, particularly among vulnerable populations such as children, the elderly, and those with pre-existing respiratory or cardiovascular conditions. The temporal analysis of air quality data from 2020 to 2024 revealed a steady increase in pollutant concentrations, with $PM_{2.5}$ rising from 92 µg/m³ in 2020 to 108 µg/m³ in 2024. Similar upward trends were observed for NO_2 and SO_2, suggesting that both vehicular emissions and industrial activities continue to drive environmental degradation. Such trends are consistent with previous findings by Basu et al. [5], who reported an association between rising $PM_{2.5}$ levels and hospital admissions in Delhi, and Campbell et al. [4], who highlighted the global health burden of fine particulate matter exposure.

The correlation heatmap provided further insight into pollutant-disease interactions. $PM_{2.5}$ demonstrated the strongest association with respiratory diseases, followed by moderate associations with circulatory system diseases and weaker but notable links with neoplasms. This is expected, as short-term exposure to fine particulate matter often results in acute respiratory conditions such as asthma or bronchitis, while long-term exposure may contribute to more chronic outcomes, including cancer. These results suggest that pollutant-specific strategies, such as stricter $PM_{2.5}$ regulations and localized interventions in high-density urban regions, could substantially reduce the health burden.

Machine learning models offered an additional layer of understanding by quantifying the predictive potential of environmental data. Gradient boosting models, namely CatBoost and XGBoost, consistently outperformed the Multilayer Perceptron (MLP) in terms of accuracy, generalization ability, and resistance to overfitting. The confusion matrices demonstrated that boosting models were better able to classify disease risks across categories, while learning curves indicated stable training and testing performance. These outcomes align with the results of Ali et al. [7], who reported the superior predictive capabilities of gradient boosting for environmental health datasets. In contrast, the MLP exhibited higher variance between training and testing accuracy, suggesting its sensitivity to relatively small, structured datasets such as those used in this study.

VII . CONCLUSION AND FUTURE WORK

This study investigated the relationship between air pollution and disease outcomes in vulnerable populations by integrating environmental datasets with health indicators. The results confirmed that $PM_{2.5}$ is the most critical pollutant, showing consistently high levels in northern Indian states and strong associations with respiratory diseases. Temporal analysis highlighted a sharp rise in pollution levels during 2021, followed by gradual declines,

reflecting the influence of changing industrial and transportation activities. Correlation analysis further established that particulate pollutants contribute significantly to respiratory health burdens, while cardiovascular and cancer outcomes show more complex patterns. The application of AI algorithms demonstrated the potential of predictive modeling for environmental health research. CatBoost and XGBoost consistently outperformed MLP, achieving high accuracy in disease prediction. These findings suggest that gradient boosting models are particularly effective for structured environmental and health datasets. Such models can support policymakers and healthcare professionals by identifying high-risk regions and forecasting disease burdens before they escalate.

Despite these contributions, the study has certain limitations. The datasets were aggregated and did not include patient-level records, which restricted the granularity of analysis. Moreover, additional pollutants such as ozone and carbon monoxide were not included, which may underestimate total pollution exposure.

Future research should expand this work by incorporating real-time IoT sensor networks, fine-grained demographic data, and deep learning approaches for more accurate disease forecasting. Integration with Geographic Information Systems (GIS) and mobile health applications could also provide community-level alerts, supporting early interventions. By bridging environmental monitoring with AI-driven health prediction, future studies can play a vital role in safeguarding vulnerable populations against the growing challenges of air pollution.

ACKNOWLEDGMENT

I sincerely thank my mentors, colleagues, and research collaborators for their guidance and support throughout this work. I acknowledge the use of Google Colab for implementing machine learning models such as MLP, CatBoost, and XGBoost, which were crucial in analyzing $PM_{2.5}$ levels, pollutant correlations, and health impacts. I appreciate the contributions of institutions and data providers, including the niti aayog for making air quality datasets available. Lastly, I am grateful to my peers, friends, and family for their encouragement, which helped me achieve the research objectives of disease prediction, correlation analysis, and environmental health assessment.

REFERENCES

1. J. Garlík and M. Trnovec, “Air pollution and its health effects: Evidence from epidemiological studies,” *Environmental Health Review*, vol. 59, no. 2, pp. 75–83, 2017.
2. P. Isola, A. Zhu, T. Zhou, and A. Efros, “Image-to-image translation with conditional adversarial networks,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2017.
3. D. Popescu and L. Ichim, “Air quality monitoring and assessment using machine learning techniques,” *International Journal of Environmental Research and Public Health*, vol. 18, no. 9, pp. 1–15, 2021.
4. J. Campbell, R. de Souza, and P. Tyrrell, “Respiratory health impacts of fine particulate matter: A global perspective,” *The Lancet Respiratory Medicine*, vol. 6, no. 10, pp. 782–790, 2018.
5. R. Basu, A. Samet, and J. Dominici, “Air pollution and hospital admissions for respiratory and cardiovascular diseases in Delhi, India,” *Journal of Epidemiology and Community Health*, vol. 73, no. 4, pp. 310–317, 2019.
6. L. Vargas-Santiago, A. Cárdenas, and J. Herrera, “Long-term cardiovascular risks of air pollution exposure: A systematic review,” *Environmental Research*, vol. 195, pp. 110–118, 2021
7. S. Ali, M. Hussain, and R. Qureshi, “Predictive modeling of air pollution and health impacts using gradient boosting methods,” *IEEE Access*, vol. 9, pp. 120453–120465, 2021.