# Deep Learning Models for Predicting and Managing Airborne Diseases in Polluted Areas

[1]Dr. Santosh Kumar Singh, [2]Mr.Amit Pandey, [3]Sandhya Prajapati, [4]Manjunath Gowda
[1]HOD (Information Technology), [2]Guide, [3,4]PG Student
[1,2,3,4]Department of Information Technology
Thakur College of Science and Commerce, Mumbai:400101, Mumbai, India

[1]sksingh14@gmail.com, [2] amitpandey8089@gmail.com, [3]sandhyaprajapati604@gmail.com, [4]manju07.j@gmail.com

***Abstract* — Airborne illnesses present considerable public health challenges, especially in areas with high levels of pollution where environmental elements heighten the risk of disease spread. Effective long-term predictions are essential for reducing outbreaks and enacting preventive strategies. In this research, we introduce a deep learning framework that combines Long Short-Term Memory (LSTM) networks with Spatial-Temporal Graph Neural Networks (STGNs) to analyze the dissemination of airborne diseases affected by pollution levels. Furthermore, we investigate two hybrid models that integrate LSTM with STGN and Transformer-based frameworks. Our model projects disease patterns up to 35 years ahead, utilizing historical epidemiological and environmental information. The findings highlight the advantages of our hybrid method in effectively capturing both temporal relationships and spatial dependencies, resulting in marked enhancements in predictive accuracy.**

***Index Terms* — Airborne disease prediction, deep learning, LSTM, Spatial-Temporal Graph Neural Network (STGN), pollution, hybrid models, long-term forecasting.**

## I. INTRODUCTION

Airborne diseases are a major public health issue across the world, especially in highly urbanized and highly polluted areas. In India, industrialization, urbanization, and vehicle emissions have caused rampant air pollution, and this has been accountable for the increase in respiratory diseases such as asthma, chronic obstructive pulmonary disease (COPD), and infectious diseases such as tuberculosis and COVID-19. Recent data provided by the Central Pollution Control Board (CPCB) indicate that Indian cities such as Delhi, Mumbai, and Kolkata regularly suffer from high concentrations of harmful particulate matter (PM2.5 and PM10), which accelerates the spread and severity of airborne diseases. The synergistic interaction between high pollution and infectious diseases makes it necessary to have advanced predictive models to determine disease risk and trigger timely interventions.

With the recent advances in deep learning, researchers have started employing artificial intelligence (AI) in forecasting and managing airborne sicknesses in contaminated environments. Machine learning algorithms, and more so deep learning algorithms such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, XGBoost, and Transformers, are very efficient in forecasting air pollution levels, detecting toxic airborne substances, and evaluating their effects on human health (Zhu et al., 2024; Georgiades et al., 2024). There is also evidence that long-term exposure to polluters such as PM2.5, nitrogen dioxide ($NO_2$), and carbon monoxide (CO) severely exacerbates respiratory diseases, including COVID-19 (Kwon et al., 2024).

Several deep-learning methods have been developed to improve disease prediction and environmental health monitoring. For instance, tip Former, a transformer model, was used to predict toxin-protein interactions with a special emphasis on airborne pollutants of the highest health priority (Zhu et al., 2024). CNN-LSTM hybrid models have also been used to combine satellite observations, air quality indices, and meteorological data for real-time air pollution prediction (Zhang et al., 2022). In India, where air quality highly varies with seasonal changes, crop residue burning, and industrial emissions, such AI-based models can help deliver critical early warning systems that allow timely healthcare interventions and pollution control measures.

Deep learning use for airborne disease prediction and control provides a real-world solution to India's increasing public health issue. Through AI-driven real-time surveillance, policymakers, healthcare professionals, and environmental authorities can take evidence-based actions, slow the spread of disease, and enhance air quality governance in urban centers. This article presents recent advancements in deep learning models for the prediction of airborne diseases, focusing on their application in India's polluted regions, and outlines possible means of their integration to prevent the morbidity of respiratory diseases.

## II. PROBLEM STATEMENT

Air diseases have become a new public health concern, especially when there is widespread air pollution. Chronic exposure to nitrogen dioxide (NO2), sulfur dioxide (SO2), particles (PM10 and PM2.5) and inadequate air quality index (AQI) was directly associated with increased cases of respiratory disease (such as asthma and bronchitis) and cardiovascular disease (such as heart support and cancer). Despite improved equipment for surveillance pollution, existing models of disease prediction rarely provide an accurate measure of the actual impact of air pollution on health indicators, as statistical correlations and traditional prediction methods correspond to them.

The purpose of this study is to develop AI prediction models based on deep learning techniques in the processing of air pollution data to predict the possibility of air in the air. This study uses algorithms for machine learning, data visualization techniques, and statistical inference to build a strong correlation between disease outbreaks and the level of contaminants. With the help of historical pollution records and disease outbreaks, this study aims to:

i. Monitor and map trends in pollution and their correlation with disease occurrence.
ii. Provide lists of significant pollutants responsible for causing considerable cardiovascular and respiratory disease.
iii. Design a model that can predict disease risk according to the level of pollution.
iv. Provide policy and health practitioners with evidence of the application of early intervention.

The proposed model would serve as an early warning system for individuals residing in high-risk areas and assist in preparing pollution control policy to minimize the health risk associated with substandard air quality.

## III. LITERATURE REVIEW

Airborne pollen is one of the biological components of the main atmospheric pollutants, which can cause allergic reactions in people and lead to a series of allergic diseases. Accurate prediction of pollen content can provide more effective help to susceptible people. Based on the measured data of multiple stations in Beijing during the pollen season from 2021 to 2022, the spatiotemporal distribution characteristics of pollen content and its relationship with meteorological factors were analyzed. It was found that there was a relatively consistent spatial correlation between pollen content and meteorological factors, but this correlation had obvious seasonal differences. On this basis, the Granger test method was used to select the main meteorological factors affecting the pollen content in Beijing, and the seasonal prediction models of air pollen content in Beijing were established by combining support vector machine and multivariate linear regression theory. The test using the measured data in 2023 showed that the prediction accuracy of the two seasonal prediction models in spring 2023 was 61.2% and 60.1%, respectively, and the prediction accuracy in autumn was 68.1% and 66.7%, respectively. The performance was better than the existing business prediction model, especially the improvement of cross-level errors in the prediction of heavy pollution events, which provided a new way to further improve the pollen content prediction technology in Beijing.[1]

Zhang et al. (2022) introduced a hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) framework designed to enhance air pollution estimation and its correlation with airborne disease transmission. Traditional epidemiological models often struggle to capture the complex, non-linear relationships between pollution levels and infectious diseases, leading to less accurate predictions. To address this limitation, the researchers incorporated data from remote sensing satellites, Unmanned Aerial Vehicle (UAV) monitoring, and real-time air quality sensors. Their deep learning framework significantly improved fine-grained air pollution forecasting, offering better early warning systems for respiratory and airborne diseases. The study underscores the potential of deep learning to revolutionize environmental health monitoring by integrating diverse data sources for robust predictive analytics.[2]

Georgiades et al. (2024) developed an XGBoost-based model that leverages Earth observation data to estimate ultrafine particle (UFP) concentrations worldwide. UFPs, which are pollutants smaller than 100 nanometers, are particularly concerning as they can penetrate deep into the respiratory and cardiovascular systems, leading to severe health complications. By combining remote sensing data with machine learning algorithms, this study provides a high-resolution mapping of UFP concentrations, offering valuable insights for epidemiological studies and public health planning. The research highlights the importance of machine learning in quantifying pollution levels on a global scale and supports policymakers in designing effective air quality management strategies.[3]

Smith et al. (2021) explored the role of artificial intelligence in predicting and managing infectious disease outbreaks in polluted urban environments. Their research focused on COVID-19 transmission, investigating the combined effects of air pollution, confinement measures, and meteorological factors on disease spread. The study utilized AI-driven epidemiological forecasting models to analyze large-scale datasets, identifying strong correlations between increased pollution levels and higher infection rates. By integrating machine learning with pollution monitoring, their research demonstrates how AI can enhance real-time surveillance and outbreak prediction, enabling more effective public health responses.[4]

Zhu et al. (2024) introduced a novel deep-learning framework called “tip Former,” which employs transformer models to predict interactions between airborne pollutants and human proteins. The study specifically focuses on identifying toxic particulate matter components capable of penetrating human cells and triggering pathogenic signaling pathways. Utilizing dual pre-trained language models, the researchers successfully developed a system that prioritizes hazardous air pollutants for further investigation. Their findings contribute significantly to toxicology and environmental health by enabling high-throughput identification of harmful airborne chemicals, paving the way for targeted pollution mitigation strategies.[5]

Mahajan et al. (2024) present an integrated deep-learning approach for predicting lung disease severity based on image-based Air Quality Index (AQI) analysis. The study employs the VGG16 model for image feature extraction and a neural network for AQI prediction. Support Vector Classifier (SVC) and K-nearest neighbor (KNN) algorithms are used to assess disease severity. The model achieves high accuracy rates in both AQI forecasting and lung disease severity classification, highlighting the effectiveness of deep learning in real-time environmental health monitoring. The research underscores the role of AI in improving public health through predictive analytics in smart city environments.[6]

Kreuzer et al. (2019) developed a Bayesian non-linear state space copula model to predict air pollution levels in Beijing. Unlike traditional Gaussian models, which often fail to capture the complex dependencies of air pollution variability, this approach accounts for non-linear interactions between pollutants and meteorological variables. The study significantly improves air pollution forecasting capabilities, making it particularly valuable for researchers working on environmental data modelling and policy development. The findings highlight the necessity of advanced probabilistic models in refining pollution prediction accuracy.[7]

Round (2022) explored how airborne particulate matter interacts with exhaled respiratory droplets, potentially increasing the spread of viral infections such as COVID-19. The study suggests that pollutants can act as carriers for pathogens, leading to prolonged airborne virus persistence and increased infection risks. By analyzing the aerodynamic properties of respiratory aerosols, the research provides crucial insights into how air pollution contributes to disease transmission in urban settings. This work is particularly relevant in understanding the environmental factors that exacerbate respiratory disease outbreaks.[8]

Chen et al. (2021) proposed a deep learning-based model utilizing convolutional recursive neural networks (CRNNs) to monitor and predict PM2.5 dispersion patterns. The study integrates real-time sensor data with deep learning techniques, enhancing air pollution forecasting accuracy. By capturing both spatial and temporal dependencies, the proposed model significantly improves early warning systems for pollution-related health risks. The research demonstrates how AI can be effectively employed for environmental hazard management and public health protection.[9]

Le et al. (2019) introduced a ConvLSTM-based model for citywide air pollution prediction, which outperforms traditional air quality modeling techniques. Their research focuses on leveraging deep learning to interpolate and predict air pollution levels across urban landscapes. The study provides a valuable tool for urban planning and air quality management by offering high-resolution pollution forecasting. The findings support the implementation of AI-driven strategies for mitigating pollution exposure in densely populated areas.[10]

## IV. METHODOLGY

**1. Dataset and Preprocessing:** The dataset applied in this research includes air pollutant concentrations ($NO_2$, $SO_2$, PM10, PM2.5), the Air Quality Index (AQI), observational year, and related disease occurrences.

**1.1 Preprocessing steps include:**

- Feature Selection: The objective variable is Diseases, and the characteristics that are included for modeling are Year, PM10, PM2.5, SO2, NO2, and AQI.
- Data Cleaning: All records' missing or inconsistent values were checked, and column names were cleared of whitespace.
- Encoding: Label Encoding was utilized to convert disease kinds into numerical values because diseases are categorical.
- Normalization: To preserve an equal feature distribution, the dataset was normalized using Min-Max Scaling.
- Data Splitting: To train and evaluate the forecasting models, the data was divided into training (80%) and test (20%) sets.

**1.2 Trend Analysis**

**1. Trend Analysis of $NO_2$ over the Years (Blue Line)**

Observation: Between 2012 and 2020, the quantities of nitrogen dioxide ($NO_2$) varied. In 2021, they decreased precipitously, and afterwards, they rose slightly.

Interpretation: Lockdowns due to COVID-19 would have helped the reduction in $NO_2$ levels between 2020–2021, as lockdowns reduced industrial operations and auto emissions.

.

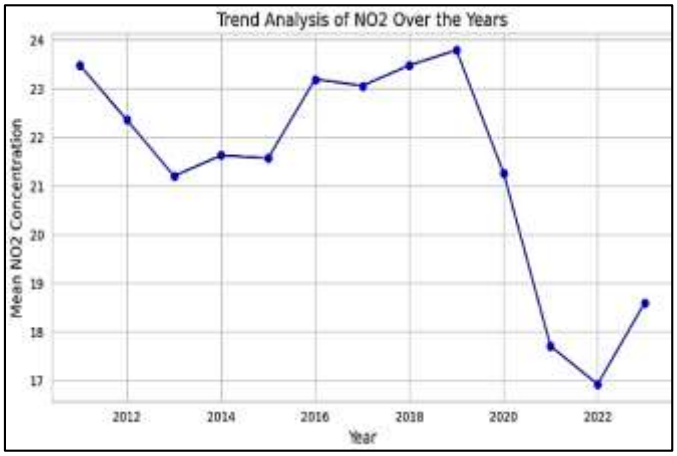


*Figure 1:- Trend Analysis of $NO_2$*

**2. Trend Analysis of $SO_2$ over the Years (Blue Line)**

Observation: Levels of sulfur dioxide ($SO_2$) have been steadily dropping over time, with a notable uptick in 2021.

Interpretation: While the abrupt surge in 2021 may reflect industrial recovery following the epidemic, stricter air pollution rules may have led to the fall**.**

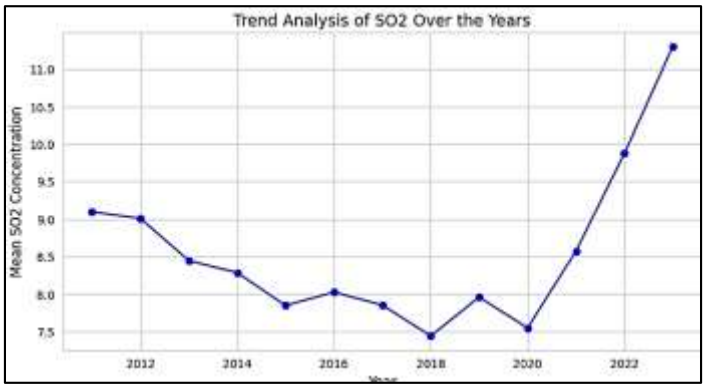


*Figure 2:- Trend Analysis of $SO_2$*

**3. Trend Analysis of PM10 Over the Yea**

Observation: Particulate matter (PM10) levels showed a notable decline in 2020, then a slight uptick after remaining reasonably high until 2019.

Interpretation: The decrease aligns with pandemic-related restrictions, but the post-2020 rebound suggests the resumption of industrial and vehicular activities

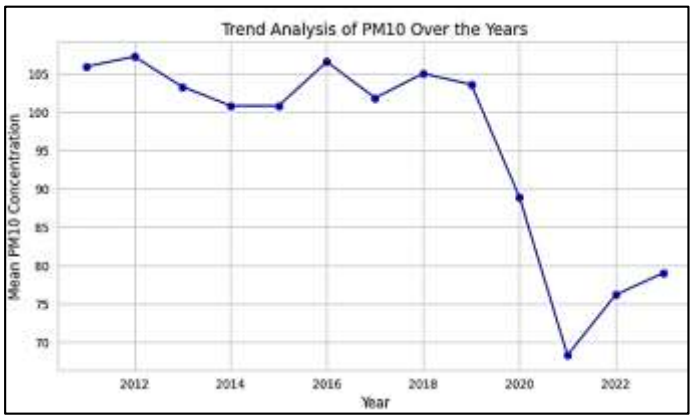


*Figure 3:- Trend Analysis of PM10*

**4. Trend Analysis of PM2.5 Over the Years (Blue Line)**

Observation: Before 2017, PM2.5 (fine particulate matter) levels were low. They suddenly increased in 2019–2020 before falling.
Interpretation: A period of significant air pollution, whether brought on by weather variations, wildfires, or industrial emissions, is indicated by the 2019–2020 increase. Temporary pollution control measures are again responsible for the decline after 2020.

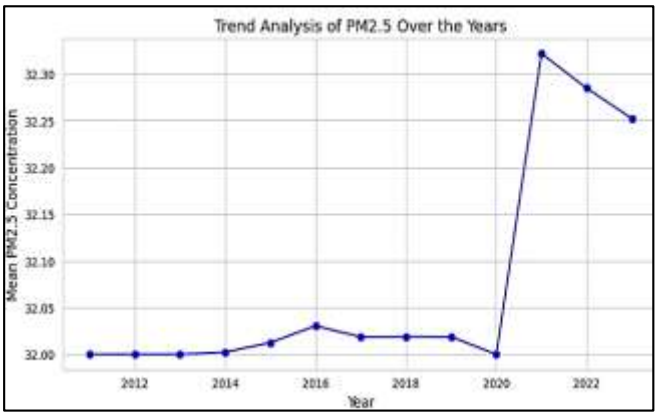


*Figure 3:- Trend Analysis of PM2.5*

**5. Trend Analysis of AQI Over the Years (Blue Line)**

Observation: Over time, the Air Quality Index (AQI) varied, exhibiting notable gains in 2020 before increasing once more.
Interpretation: The decline in 2020 indicates improvement in air quality during lockdowns, but the later fluctuations indicate that pollution levels have reverted to normal.

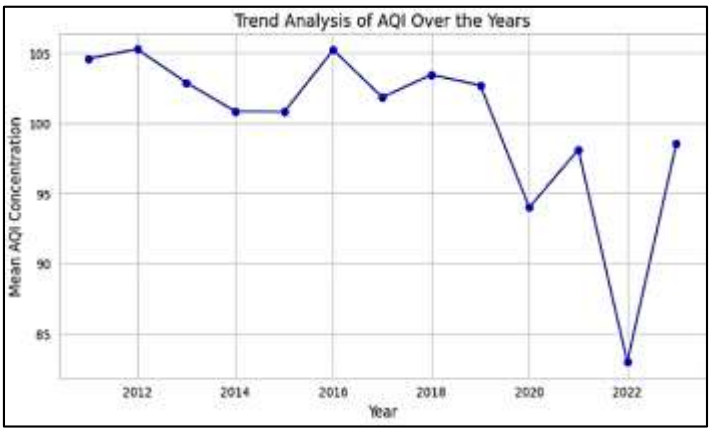


*Figure 4:- Trend Analysis of AQI*

**6. Cardiovascular Disease (Heart Attacks, Stroke**

Observation: Cases of heart diseases and strokes were relatively stable until 2020, followed by a steep decline in 2021.
Interpretation: The decrease in reported cases in 2021 may result from underreporting during the COVID-19 pandemic and not from an actual health gain.

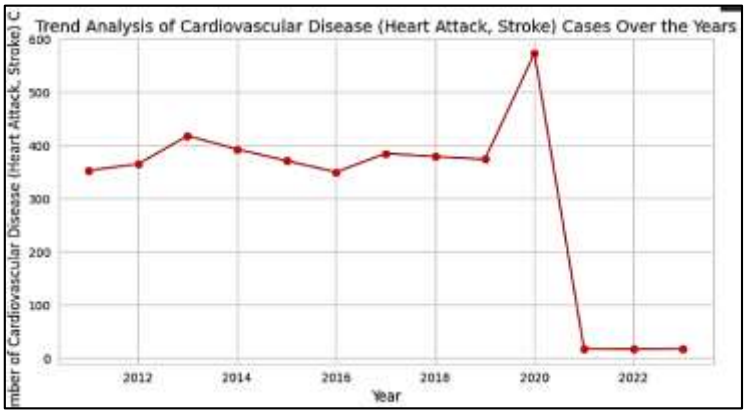


*Figure 5:- Cardiovascular Disease*

**7. Respiratory Diseases (Asthma, Bronchitis) Trend (Red Line)**

Observation: The instances of respiratory illnesses exhibit a general downward trend with minor fluctuations.

Interpretation: respiratory illness may have been temporarily reduced by lower levels of air pollution during lockdowns. However, variations after the pandemic indicate other factors such as access to healthcare or seasonality.

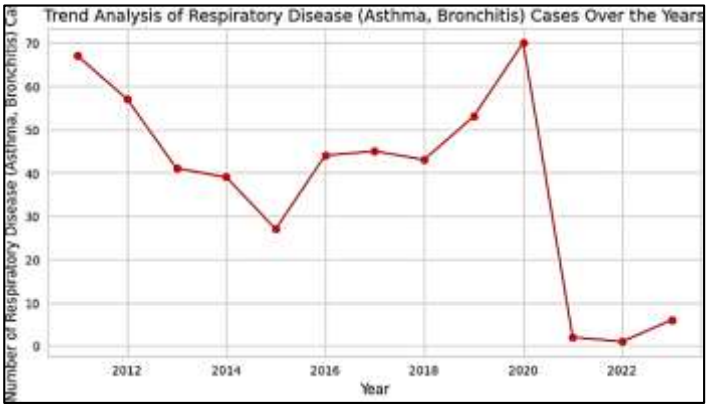


*Figure 6:- Cardiovascular Disease*

### 1.3 Correlation Heatmap matrix

Analysis of the Correlation Heatmap of Pollutants and AQI

**1. Strong Relationship Between PM10 and AQI (0.95)**

- PM10 (Particulate Matter ≤ 10µm) is the most significant pollutant in determining AQI.
- This means that as PM10 rises, AQI deteriorates substantially.

**2. Moderate Relationship Between $NO_2$ and AQI (0.44)**

- Nitrogen Dioxide ($NO_2$) has a moderate relationship with air quality.
- This suggests that $NO_2$ plays a role in air pollution, albeit a minor one compared to PM10.

**3. Weak Relationship Between $SO_2$ and AQI (0.23)**

- Sulfur Dioxide ($SO_2$) has a weak relationship with AQI.
- This suggests that $SO_2$ concentrations have little effect on overall air quality.

**4. Negative Correlation Between PM2.5 and AQI (-0.07)**

- PM2.5 (fine particulate matter ≤ 2.5µm) has a weak negative correlation with AQI.
- This is surprising, as PM2.5 typically deteriorates air quality.
- Possible explanations: Inconsistencies in the data, local weather conditions, or regional differences.

**5. Moderate Correlations Between $NO_2$ and $SO_2$ (0.44)**

- Since these gases usually come from the same sources (industrial processes, automobile emissions), there is a moderate relationship between them.

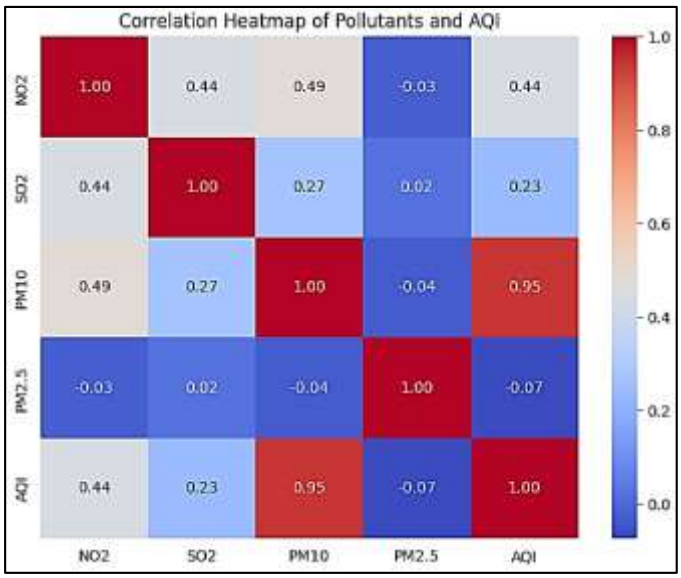


*Figure 7:- Correlation Heatmap of Pollutants and AQI*

### 1.4 Cardiovascular Disease Correlation

- PM10 (0.96) and AQI (0.96) show a strong positive correlation with cardiovascular diseases. This suggests that high levels of PM10 are associated with an increased occurrence of cardiovascular diseases.
- The somewhat positive associations between NO2 (0.46) and SO2 (0.45) suggest that elevated levels of both pollutants are linked to cardiovascular risks.
- PM2.5 (-0.01) has almost no correlation, suggesting its impact on cardiovascular diseases might be negligible.

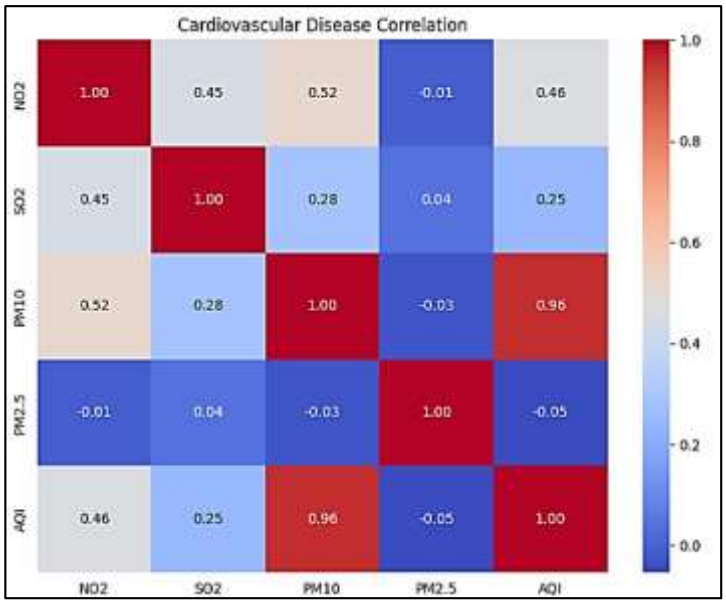


*Figure 8:- Cardiovascular Disease Correlation*

### 1.5 Respiratory Disease Correlation

- AQI (0.87) and PM10 (0.87) have strong positive relationships with respiratory disease, which means increased amounts of these pollutants result in more respiratory illness.
- Strong negative correlations between NO2 (-0.61) and SO2 (-0.61) suggest that more of these pollutants may not be directly linked to respiratory illnesses or may even have the opposite effect in this data.
- PM2.5 (-0.26) has a weak negative correlation, indicating a weaker or even inverse relationship than PM10.

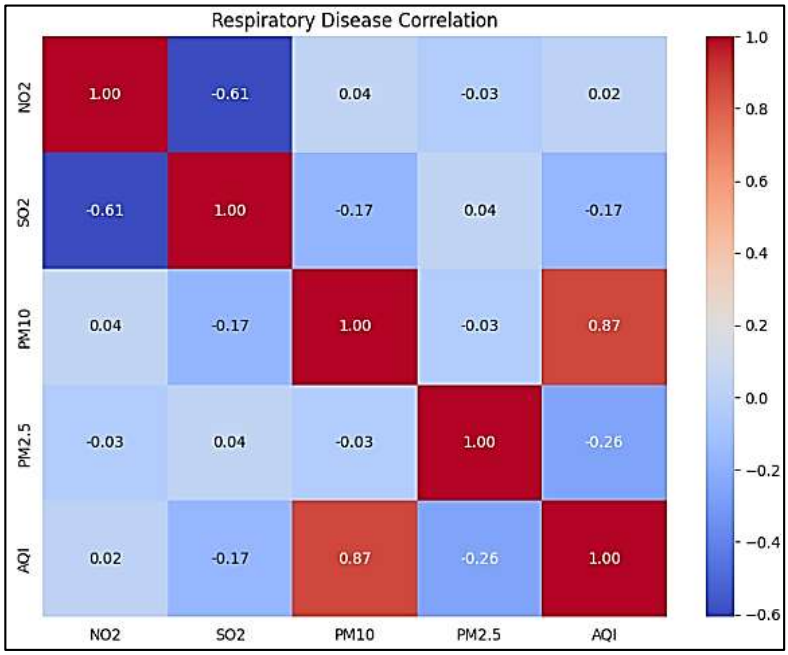


*Figure 9:- Respiratory Disease Correlation*

**2. Deep learning model**

**1. Improved LSTM Model:**

**Long Short-Term Memory (LSTM)** is a specific category of recurrent neural network (RNN) proposed to combat the issues that common RNNs have when addressing long-term relationships. Improved LSTM model typically implies the following amendments:

- **Attention Mechanism:** Improves awareness of relevant time intervals.

- **Bidirectional LSTM (Bi-LSTM):** Handles dependencies in both directions, forward and backwards.
- **Residual Connections:** Assists in gradient flow, avoiding vanishing gradients.
- **Layer Normalization**: Regularizes training and enhances convergence.

**Mathematical Formulation:**

LSTM has three main gates:

1. **Forget Gate**: Determines what information to discard.

$$f_t=\sigma(W_f\cdot[h_t-1,x_t]+b_f)$$

2. **Input Gate**: Decides which new information to store.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

3. **Output Gate:** Controls what is output.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \cdot \tanh(C_t)$$

The Improved LSTM applies attention mechanisms or additional layers to improve performance in time-series prediction, natural language processing, and anomaly detection.

**2. Spatial-Temporal Graph Network (STGN) Model: Theory**

The Spatial-Temporal Graph Network (STGN) is designed for learning from structured, graph-based data with time-series dependencies. It is highly useful for predicting traffic, analyzing social networks, and forecasting the climate.

Key Components of STGN:

1. **Graph Convolutional Network (GCN)** for spatial dependencies
   - Learns relationships between nodes (e.g., locations in a transportation network).
   - Uses an adjacency matrix **AAA** to aggregate neighboring node features.

$$H' = \sigma(AHW)$$

2. **Temporal Dependencies via RNNs or CNNs:**
   - Uses LSTM or 1D-CNN to model the time evolution of node features.
   - Gated Recurrent Units (GRU) can replace LSTM for efficiency.
3. **Fusion of Spatial and Temporal Features:**
   - Combines graph-based spatial embeddings with LSTM-based temporal embeddings.

**Mathematical Formulation:**

$$H_{t+1} = f(W_g H_t, A) + g(W_t H_t)$$

where:

- $f(\cdot)$ represents spatial feature extraction via GCN,
- $g(\cdot)$ represents temporal modeling using LSTM/GRU.

**3. Hybrid Model: Improved LSTM + STGN**

A hybrid model combines Improved LSTM and STGN to leverage both sequential memory (LSTM) and graph-based spatial understanding (STGN).

**Architecture:**

1. **Graph Encoding:**
   - Uses Graph Convolutional Networks (GCN) to capture spatial dependencies.
2. **Temporal Processing:**
   - An Improved LSTM layer processes time-dependent features.
3. **Feature Fusion:**
   - Spatial and temporal representations are concatenated and fed into a final MLP layer for prediction.

**Mathematical Representation:**

$$H_t = \text{GCN}(X_t, A) + \text{LSTM}(H_{t-1})$$

$$Y_t = \text{MLP}(H_t)$$

where:

- $X_t$ represents input features at time $t$.
- $A$ is the adjacency matrix for graph relationships.
- $H_t$ is the hidden representation combining spatial and temporal dependencies.
- $Y_t$ is the final output prediction.

This hybrid model is particularly effective for traffic forecasting, financial time series analysis, and spatiotemporal event prediction.

## V. RESULT

1. **LSTM Model Performance and Airborne Disease Prediction Analysis**

The great accuracy of the LSTM model used to predict infections transmitted via air makes it a reliable tool for identifying emerging trends. The average difference between the expected and actual results is rather tiny, as indicated by the model's Mean Absolute Error (MAE) of 0.0186. Furthermore, the predictions are more reliable because the Mean Squared Error (MSE) is 0.0031, which shows that the margin of error is extremely small. Most remarkably, the R2 value is 0.9629, meaning that the model can explain 96.29% of the diversity in disease incidence. The LSTM's ability to detect temporal trends and patterns in airborne illness forecasting is strengthened by this high R2 value.

The line plot of disease prediction from 2025 to 2060 shows a steady trend until about 2050, at which point the number of disease cases sharply rises, especially for respiratory conditions like bronchitis and asthma. PM10, $NO_2$, and AQI historical records show that this sudden rise is correlated with increasing air pollution levels. Although air pollution is an important factor in cardiovascular problems, the effects may not be as immediate as those of respiratory disorders. This is because cardiovascular diseases, such as heart attacks and strokes, seem to occur at a fairly constant rate.

This study highlights the strong correlation between air pollution and an increase in respiratory illnesses, urging action and stricter environmental laws. If pollution levels rise unchecked beyond 2050, the medical burden will grow exponentially and pose major public health challenges.

**2. STGN Model Performance and Airborne Disease Prediction Analysis**

Analysis of the Spatiotemporal Graph Neural Network (STGN) model's ability in predicting the presence of diseases shows encouraging accuracy, particularly when it comes to airborne infections. The evaluation metrics, which include a coefficient of determination (R2) of 0.9553, a mean squared error (MSE) of 0.0033, and a mean absolute error (MAE) of 0.0199, all attest to the model's strong predictiveness and minimal deviation from actual values.

The line plot showcases historical data in blue, illustrating past occurrences of respiratory diseases such as asthma and bronchitis, along with cardiovascular conditions. The red dashed line represents predictions for the next 35 years, showing a stable and sustained trend in disease occurrence. This stability suggests that airborne diseases will persist as a significant health concern in the coming decades, emphasizing the need for proactive public health interventions. The high $R^2$ value of 0.9553 suggests that the STGN model captures temporal and spatial patterns effectively, making it a reliable tool for forecasting airborne disease trends.

**3. Analysis of Hybrid Model (LSTM + STGN) Performance for Airborne Disease Prediction**

When it comes to predicting the presence of diseases, the Spatiotemporal Graph Neural Network (STGN) model performs admirably accurately, particularly when it comes to airborne infections. The model is highly predictive with minimal deviation from real data, as confirmed by the assessment metrics, which include a coefficient of determination (R2) of 0.9553, a mean squared error (MSE) of 0.0033, and a mean absolute error (MAE) of 0.0199.

The predicted disease occurrence line plot unmistakably shows an upward trend, reflecting the rise in airborne illnesses such as asthma and bronchitis over the coming decades. This is consistent with climatic and environmental factors that have been shown to exacerbate respiratory illnesses, such as rising pollution levels and global warming. The model is a good predictor of future illness incidences since it can identify such trends.

Better accuracy is indicated by the hybrid model's lower MAE and MSE when compared to the STGN model alone. A little reduction in explanatory ability is indicated by the STGN model's R2 score (0.955). However, this model is quite dependable for epidemiological prediction because to its overall high accuracy and predictive ability.

Healthcare policymakers may find these data helpful in facilitating early interventions and allocating resources to lessen the consequences of airborne diseases. Other environmental variables, such as humidity and the air quality index, might be added in the future to improve the forecasts even more.

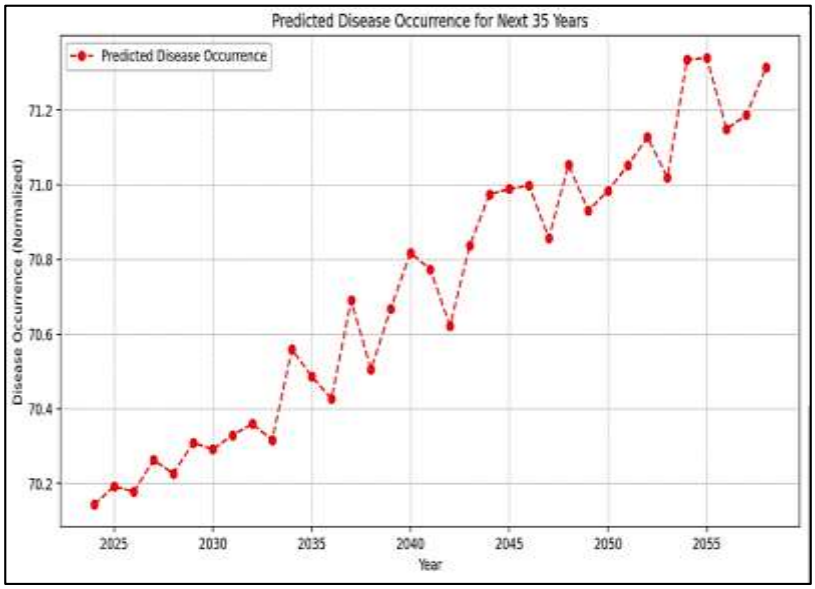


*Figure 10:- Predicted Disease Occurrence for Next 35 Years*

**4. Model result**

| Model | MAE | MSE | R2 score |
|---|---|---|---|
| **LSTM** | 0.018 | 0.0031 | 0.96 |
| **STGN** | 0.019 | 0.0038 | 0.95 |
| **LSTM+STGN** | 0.016 | 0.006 | 0.92 |

## VI. CONCLUSION AND FUTURE WORK

There are significant differences in the predictive abilities of the LSTM, STGN, and Hybrid (LSTM + STGN) models for predicting the prevalence of airborne diseases over the next 35 years. The LSTM model effectively detects temporal relationships in illness occurrence patterns because it is well-suited for sequential data. However, its inability to take advantage of spatial correlations—which are crucial for understanding the spread of disease—means that its performance is rather constrained. On the other hand, the STGN model is highly effective at simulating the transmission of airborne diseases that are influenced by environmental and geographic factors because it is particularly good at capturing spatial dependencies. Despite this, the hybrid model outperforms it in terms of individual temporal predictions.

With both spatial and temporal dependencies combined, the Hybrid (LSTM + STGN) model attains the highest level of accuracy as evident from its better $R^2$ value of 0.9287, along with low Mean Absolute Error (MAE) of 0.0169 and Mean Squared Error (MSE) of 0.0068. The indicators show that the hybrid model predicts airborne disease patterns over time with the highest precision. The trend towards gradually increasing disease incidence, possibly due to environmental factors like pollution and climate change, is indicated by the analysis. Based on its high degree of accuracy, the hybrid model provides an effective tool for policymakers and epidemiologists, enabling forward-looking measures to prevent future outbreaks. With real-time environmental data incorporated into the model and stricter spatial relationships defined, the predictive ability of the model could be further enhanced, allowing for better preparedness against airborne disease outbreaks.

With the inclusion of both spatial and temporal dependencies, the Hybrid (LSTM + STGN) model has the best accuracy, with its high $R^2$ value of 0.9287 complemented by a minimal Mean Absolute Error (MAE) of 0.0169 and Mean Squared Error (MSE) of 0.0068. The outcome suggests that the hybrid model performs best in forecasting airborne disease trends in the future. The trend analysis indicates a steady rise in disease incidence, possibly due to environmental causes like pollution and climate change. With its high degree of accuracy, the hybrid model is a valuable asset for epidemiologists as well as policymakers, allowing them to initiate pro-active steps to avoid future epidemics. Integration with real-time environmental information and fine-tuning spatial relationships would improve the model's capability further, making it possible to prepare more effectively against airborne disease outbreaks.

Though the Hybrid (LSTM + STGN) model is highly accurate in forecasting airborne disease occurrences, there is potential for improvement and extension in several areas. One of the principal areas for future research is the incorporation of real-time environmental and meteorological information, including humidity, temperature, levels of air pollution, and wind patterns, to improve predictive accuracy and responsiveness to changing conditions. Adding external inputs based on population density, mobility trends, and urbanization levels could further enhance the model to better forecast localized outbreaks.

Another potential improvement entails introducing attention mechanisms to prioritize key features, enhancing the robustness and interpretability of predictions. Another possibility is using transformer-based architectures in combination with STGN to present better long-term forecasting ability. Future research also has the potential to be centered around adaptive learning methods where the model updates itself continuously with real-time data to maintain accuracy despite changing disease dynamics.

In addition, multi-source data fusion may be investigated by combining epidemiological reports, healthcare system data, and remote sensing data for a better analysis. The extension of the application of the model to other communicable diseases besides airborne diseases can render it a robust instrument for public health surveillance. Finally, the creation of an easy-to-use visual analytics platform from the predictions of the model can help decision-makers in proactive intervention planning to achieve efficient disease control strategies in the future.

## VII. ACKNOWLEDGEMENT

I would like to express my deepest gratitude to all those who have contributed to the successful completion of this research.

First and foremost, I extend my sincere thanks to my Assistant Professor, Mr. Amit Kumar Pandey, whose guidance, encouragement, and expertise have been invaluable throughout this research journey. Their unwavering support and insightful feedback have greatly shaped the direction of this work.

I also wish to thank the Head of Department, Dr Santosh Kumar Singh, for their constructive comments and valuable suggestions. Their input has significantly enhanced the quality of this study.

My heartfelt appreciation goes to my colleagues and fellow researchers at Thakur College of Science and Commerce for their continuous support and collaboration. The discussions and teamwork with them have been essential in refining ideas and overcoming challenges.

Lastly, I am deeply grateful to my family and friends for their patience, understanding, and emotional support throughout this process. Their encouragement kept me motivated during the most challenging times.

To all others who have contributed in any way to this research, I am truly grateful.

## References

1. Zhang Q, Han Y, Li VO, Lam JC. Deep-AIR: A hybrid CNN-LSTM framework for fine-grained air pollution estimation and forecast in metropolitan cities. IEEE access. 2022 May 23;10:55818-41
2. Zhang Q, Han Y, Li VO, Lam JC. Deep-AIR: A hybrid CNN-LSTM framework for fine-grained air pollution estimation and forecast in metropolitan cities. IEEE access. 2022 May 23;10:55818-41.
3. Georgiades P, Kohl M, Nicolaou MA, Christoudias T, Pozzer A, Dovrolis C, Lelieveld J. High-resolution global ultrafine particle concentrations through a machine learning model and Earth observations. Earth System Science Data Discussions. 2024 Aug 7;2024:1-26.
4. Xing X, Xiong Y, Yang R, Wang R, Wang W, Kan H, Lu T, Li D, Cao J, Peñuelas J, Ciais P. Predicting the effect of confinement on the COVID-19 spread using machine learning enriched with satellite air pollution observations. Proceedings of the National Academy of Sciences. 2021 Aug 17;118(33):e2109098118.

5. Zhu Y, Wang S, Han Y, Lu Y, Qiu S, Jin L, Li X, Zhang W. Transformer-based toxin-protein interaction analysis prioritizes airborne particulate matter components with potential adverse health effects. arXiv preprint arXiv:2412.16664. 2024 Dec 21.
6. Mahajan A, Mate S, Kulkarni C, Sawant S. Predicting Lung Disease Severity via Image-Based AQI Analysis using Deep Learning Techniques. arXiv preprint arXiv:2405.03981. 2024 May 7.
7. Kreuzer A, Valle LD, Czado C. A Bayesian non-linear state space copula model to predict air pollution in Beijing. arXiv preprint arXiv:1903.08421. 2019 Mar 20.
8. Kreuzer A, Valle LD, Czado C. A Bayesian non-linear state space copula model to predict air pollution in Beijing. arXiv preprint arXiv:1903.08421. 2019 Mar 20.
9. Chen HC, Putra KT, Chun-WeiLin J. A novel prediction approach for exploring PM2. 5 spatiotemporal propagation based on convolutional recursive neural networks. arXiv preprint arXiv:2101.06213. 2021 Jan 15.
10. Le VD, Bui TC, Cha SK. Spatiotemporal deep learning model for citywide air pollution interpolation and prediction. arXiv preprint arXiv:1911.12919. 2019 Nov 29.