Correlation Between Air Quality And Respiratory Health

BY

Chinmoy Saha
ID: 242-25-041

This Report Presented in Partial Fulfillment of the Requirements for
The Degree of Masters of Science in Computer Science and Engineering

Supervised By

Dr. S M. Aminul Haque
Professor And Associate Head

Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

AUGUST 2025

APPROVAL

This Project/Thesis titled **"Correlation Between Air Quality And Respiratory Health"** submitted by **"Chinmoy Saha"** ID No: **"242-25-041"** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13-09-2025.

BOARD OF EXAMINERS

Chairman

Dr. Sheak Rashed Haider Noori
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Md. Zahid Hasan
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Naznin Sultana
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Mr. Nazibur Rahman
Head of IT Infrastructure
Networld Bangladesh PLC

DECLARATION

I hereby declare that, this project has been done by me under the supervision of **Dr. S. M. Aminul Haque, Professor And Associate Head, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. S. M. Aminul Haque
Professor And Associate Head
Department of CSE
Daffodil International University

Submitted by:
Chinmoy
Chinmoy Saha
ID: 242-25-041
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I would like to express my heartiest thankfulness and gratefulness to almighty God for the divine blessing that made me capable of completing the final year thesis successfully. I am really grateful and wish my profound indebtedness to my dear supervisor Dr. S. M. Aminul Haque, Professor and Associate Head, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of "MACHINE LEARNING" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project. I would like to express my heartiest gratitude to and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University. I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work. Finally, I must acknowledge with due respect the constant support and patients of my parents.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

This thesis analyzes the relationship between air quality and respiratory health that develops a predictive framework to inform public health planning. Measurements of key pollutants (PM2.5, PM10, NO2, SO2, O3) and meteorological variables (temperature, humidity, wind) were combined with respiratory health outcomes to quantify associations and forecast short-term risk. After data preparation and normalization, an artificial neural network (ANN) was trained for regression (estimating health burden) and classification (assigning risk categories) and evaluated using standard metrics. The analysis indicates that elevated concentrations of fine particulates and nitrogen dioxide are consistently associated with increased respiratory morbidity, while meteorological conditions modify exposure–response patterns and improve predictive performance. The resulting model demonstrates practical utility for early warnings, clinical preparedness, and targeted mitigation in high-risk locations. Overall, integrating environmental monitoring with machine learning can shift practice from reactive management to proactive prevention. Future work should evaluate generalizability across regions, incorporate additional health endpoints and socioeconomic context, and assess operational deployment and cost-effectiveness under changing climate conditions.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
CHAPTER	
CHAPTER 1: INTRODUCTION	**1-5**
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	2
1.4 Research Questions:	3
1.5 Expected Output	3
1.6 Project Management and Finance	4-5
CHAPTER 2: BACKGROUND	**6-9**
2.1 Preliminaries/Terminologies	6-7
2.2 Related Works	7
2.3 Comparative Analysis and Summary	7-8
2.4 Scope of the Problem	8
2.5 Challenges	9
CHAPTER 3: RESEARCH METHODOLOGY	**10-12**
3.1 Research Subject and Instrumentation	10
3.2 Data Collection Procedure / Dataset Description	10
3.3 Statistical Analysis	11
3.4 Methodology / Analytical Framework	11
3.5 Implementation Requirements	12
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	13-24
4.1 Experimental Setup:	13-14
4.2 Experiment results and analysis:	15-23
4.3 Discussion:	24
CHAPTER 5: CONCLUSION AND FUTURE WORK	**25-26**

5.1 Impact on Society	25
5.2 Impact on Environment	25
5.3 Ethical Aspects	25-26
5.4 Sustainability Plan	26
CHAPTER 6: CONCLUSION AND FUTURE WORK	**27-28**
6.1 Summary of the Study	27
6.2 Conclusions	27
6.3 Implication for Further Study	28
REFERENCES	**29-30**

LIST OF FIGURES

FIGURES	PAGE NO
Figure 4.2.1 Box Plots	15
Figure 4.2.2 Density plots / distributions	16
4.2.3 Correlation heat-map	17
4.2.4 Original vs transformed features	18
4.2.5 Model loss	19
4.2.6 Class distribution — before merging	20
4.2.7 Class distribution — after merging	21
4.2.8 Model loss — final run	22
4.2.9 Model accuracy — final run	23

CHAPTER 1
INTRODUCTION

1.1 Introduction:

Bangladesh is currently grappling with a major environmental health challenge caused by extremely poor air quality (Air Quality Life Index [AQLI], 2022; World Bank & DoE, 2019) [2, 7]. The country ranks among the highest in the world for air pollution, with average yearly levels of fine particulate matter (PM2.5) far above WHO safety limits set by World Health Organization, 2013) [1]. This air pollution has been strongly linked to a growing number of long-term respiratory illnesses. (Greenstone, Hasenkopf, & Lee, 2022) [10]. Diseases such as asthma, COPD and serious lung infections are becoming more common, especially among children, older adults, and people living in packed urban areas like Dhaka. Pollution in Bangladesh comes from both outside and inside the home. (World Bank & DoE, 2019) [7]. Outdoor sources include exhaust from vehicles, emissions from factories, and dust from ongoing construction. Indoors, many families still rely on burning wood, dung, or crop waste for cooking and heating, which adds to the problem. The impact on health is severe. Air pollution is contributing to early deaths, lower life expectancy, and widespread illness. There is an urgent need for more detailed studies to understand how different levels of exposure affect health, how these effects change over time, and the specific ways pollution harms the lungs. This information is essential for shaping effective health policies and solutions in Bangladesh. The research analyzes how air quality factors relate to and predict respiratory conditions in Bangladesh, based on combined daily environmental data and hospital case records

1.2 Motivation:

This thesis investigates the correlation between air quality and respiratory health using an ANN-centered analytical framework that integrates classification and regression tasks relevant to environmental and clinical decision-making. The accompanying codebase implements end-to-end workflows for data ingestion, cleaning, normalization, model training, and evaluation, enabling both prediction of air quality indicators and categorization of exposure-related risk. By focusing on model architectures capable of capturing non-linearity and interactions among pollutants and meteorological covariates, the work aims to (i) quantify the strength and form of associations between pollutant profiles and respiratory risk proxies, (ii) assess the incremental value of ANN models relative to simpler baselines, and (iii) explore how feature importance and partial dependence vary across seasons and pollutant regimes.

1.3 Rationale of the Study:

Air pollution remains one of the most significant modifiable determinants of respiratory morbidity. While many studies document associations between pollutants and health, decision-makers still need practical tools that translate monitoring data into timely, local risk information. This study addresses that gap by integrating multi-pollutant and meteorological data to model respiratory risk with an artificial neural network. The approach serves two purposes: (i) quantify short-term relationships between ambient pollutants and respiratory outcomes, and (ii) generate operational forecasts to guide alerts, clinic preparedness, and targeted interventions. The work is also motivated by equity concerns, as exposure and health burdens are unevenly distributed across communities. By producing interpretable, location-ready predictions, the study aims to complement surveillance systems and inform policies that reduce preventable exacerbations and associated costs.

1.4 Research Questions:

Air pollution remains one of the most significant modifiable determinants of respiratory morbidity. While many studies document associations between pollutants and health, decision-makers still need practical tools that translate monitoring data into timely, local risk information. This study addresses that gap by integrating multi-pollutant and meteorological data to model respiratory risk with an artificial neural network. The approach serves two purposes: (i) quantify short-term relationships between ambient pollutants and respiratory outcomes, and (ii) generate operational forecasts to guide alerts, clinic preparedness, and targeted interventions. The work is also motivated by equity concerns, as exposure and health burdens are unevenly distributed across communities. By producing interpretable, location-ready predictions, the study aims to complement surveillance systems and inform policies that reduce preventable exacerbations and associated costs.

1.5 Expected Output:

- To what extent are short-term variations in ambient pollutants (e.g., PM2.5, PM10, NO2, SO2, O3) associated with changes in respiratory health outcomes?
- Does incorporating meteorological factors (e.g., temperature, humidity, wind) improve the prediction of respiratory risk relative to pollutant-only models?
- Can an artificial neural network accurately estimate daily respiratory burden (regression) and classify risk levels (classification) suitable for operational use?
- Which pollutants and meteorological features most strongly influence predicted risk, and how stable are these influences over time?
- How can model outputs be translated into actionable insights for healthcare planning, public advisories, and protection of vulnerable populations?

1.6 Project Management and Finance:

- A cleaned and harmonized dataset linking air pollutants, weather variables, and respiratory health outcomes at daily or sub-daily resolution.
- Predictive modeling
- Regression model to estimate daily respiratory burden (with performance metrics such as R^2, MAE/RMSE).
- Classification model to categorize risk levels (with accuracy, precision/recall, F1, and AUC).
- Feature influence analysis (e.g., sensitivity or SHAP-style summaries) highlighting key drivers and their relative contributions.
- An operational concept for early warning: threshold definitions, lead times, and user-oriented outputs (e.g., risk categories and suggested actions).
- Documentation of methodology, validation strategy (including time-aware evaluation), and guidelines for updating and monitoring model drift.
- Policy- and practice-oriented recommendations for targeted mitigation and communication.
- Work-plan and Timeline
- Phase 1: Scoping and data access (weeks 1–3) — define study area and period; secure pollutant, meteorology, and health data; finalize variables.
- Phase 2: Data engineering (weeks 4–7) — cleaning, alignment, feature construction, and exploratory analysis.
- Phase 3: Modeling (weeks 8–12) — train/validate ANN for regression and classification; benchmark against baseline models; tune hyper parameters.
- Phase 4: Explainability and evaluation (weeks 13–15) — feature influence, temporal generalization tests, and robustness checks.
- Phase 5: Operationalization concept (weeks 16–18) — define thresholds, alert logic, and stakeholder-oriented outputs.
- Phase 6: Writing and dissemination (weeks 19–22) — compile results, draft thesis chapters, and prepare presentation materials.
- Roles and Responsibilities

- Principal researcher: data handling, modeling, analysis, and documentation.
- Supervisor/committee: methodological guidance, review, and ethical oversight.
- Optional collaborators: domain experts (public health, meteorology) for interpretation and validation.
- Risk Management
- Data gaps: predefine imputation strategies and backup data sources; document sensitivity to missing
- Model overfitting: use time-aware validation, regularization, and out-of-sample tests.
- Drift and generalizability: reserve a holdout period; propose a schedule for periodic retraining.
- Ethical/compliance: ensure privacy safeguards for health data and transparent reporting of uncertainty.
- Budget and Resources
- Computing: access to a workstation or cloud credits for model training and storage.
- Software: open-source tools (Python/R) for preprocessing, modeling, and visualization.
- Data: public monitoring and meteorological feeds; any fees for proprietary health datasets (if applicable).
- Contingencies: modest allocation for sensor calibration, data acquisition, or conference dissemination if relevant.Note: If your program requires numeric budgets, insert estimated costs for data access (if any), cloud compute, and publication/dissemination.

CHAPTER 2
BACKGROUND

2.1 Preliminaries/Terminologies

- Air Quality: The state of outdoor air determined by the presence and levels of contaminants relative to health-based guidelines.
- Air Quality Index (AQI): A composite scale that translates pollutant concentrations into categories of health concern for public communication.
- Particulate Matter (PM):
- PM2.5: Fine particles ≤2.5 μm that reach deep lung regions and are strongly linked to respiratory and cardiovascular harm.
- PM10: Coarser particles ≤10 μm that mainly affect upper airways and trigger symptoms.
- Ozone (O3): A secondary oxidant formed from nitrogen oxides and volatile organics under sunlight; associated with airway inflammation and reduced lung function.
- Nitrogen Dioxide (NO2): A marker of combustion, especially traffic; irritates airways and is tied to asthma exacerbations.
- Sulfur Dioxide (SO2): Produced by burning sulfur-containing fuels and some industrial processes; can induce broncho constriction and symptoms in sensitive people.
- Meteorological Modifiers: Temperature, humidity, wind, and boundary layer dynamics that shape pollutant formation, dispersion, and persistence.
- Respiratory Health Outcomes: Measures such as daily respiratory clinic visits, emergency department attendances, hospital admissions, medication use, and disease-specific counts (e.g., asthma, COPD).
- Exposure Metrics: Temporal aggregations (hourly, daily, multi-day), lag structures, and cumulative indicators that capture acute and short-term effects.
- Modeling Approaches:

- Machine Learning (e.g., ANN, gradient boosting, random forest): Optimized for prediction and capturing complex, non-linear patterns.
- Statistical/Econometric Models (e.g., GLM, GAM, DLNM): Suited for estimating effect sizes, uncertainty, and testing hypotheses.
- Risk Thresholds: Cut-points used to define health risk levels for alerts and interventions, often aligned with AQI categories or outcome-derived thresholds.

2.2 Related Works

Research from Bangladesh and other developing nations repeatedly shows that poor air quality is closely linked to increased rates of respiratory illness. (Darain et al., 2013; Bahauddin & Uddin, 2010) [4, 12]. In Dhaka, for example, studies have revealed that when fine particulate matter (PM2.5) levels rise, hospital visits for breathing problems also increase. (Yeasmin et al., 2021; Sherris et al., 2021) [14, 15]. Emissions from dieselpowered vehicles and brick kilns are particularly damaging to respiratory health. (Darain et al., 2013) [4]. Other findings also highlight traffic and industrial pollution as key causes of long-term breathing issues in urban areas. (EDGAR, 2022) [8] According to the World Health Organization, outdoor air pollution ranks among the world's leading health threats, contributing to millions of early deaths annually. (World Health Organization, 2013) [1]. Children and older adults face the greatest health risks from exposure. (Sherris et al., 2021; Bontinck et al., 2020) [15, 13]. While most earlier investigations used standard statistical methods to study the problem, newer research increasingly uses machine learning for more accurate predictions. However, in Bangladesh, very few studies combine both air quality measurements and health data, leaving an important gap in existing research.

2.3 Comparative Analysis and Summary

- Modeling Paradigms:
- Statistical (GAM/DLNM): Clear effect estimates and uncertainty; may struggle with highly non-linear, high-dimensional interactions.

- ML/Deep Learning (RF/GBM/ANN/LSTM): Strong predictive accuracy; requires additional tools for interpretability and does not, by itself, establish causality.
- Exposure Characterization:
- Monitor-only data provide robust temporal signals but limited spatial coverage.
- Enhanced approaches (LUR, satellite, CTM, sensors) improve spatial granularity but need calibration and can introduce modeling bias.
- Outcome Data:
- Hospital and ED records are clinically robust but less timely and sometimes restricted.
- Syndromic data and medication use offer faster signals with more noise.
- Meteorological Context:
- Including weather and seasonality substantially improves estimates and forecasts; omission reduces skill and can bias associations.
- Deployment Considerations:
- For operations, categorical risk levels, lead time optimization, and probabilistic outputs are valuable.
- Time-aware validation (e.g., rolling-origin) better reflects real-world performance than random splits.

Summary: The literature consistently links degraded air quality with worse respiratory outcomes. Statistical models excel at inference; machine learning often leads in prediction. (Boogaard et al., 2019) [3]. Best practice combines improved exposure estimates, meteorological context, rigorous validation, and interpretable outputs to support decisions.

2.4 Scope of the Problem

Bangladesh's urban areas are experiencing significant deterioration in air quality, driven by rapid industrial growth, accelerated urban expansion, and rising emissions from motor vehicles. (DoE, 2020) [6]. Air pollution carries immediate health risks and broader environmental consequences (Gurjar, Molina, & Ojha, 2010) [11]. Civil citizens often notice worsening air pollutions linking it both to industrial and household works. (Environment and Social Development Organization [ESDO], 2020) [9]. A major barrier

 to accurately assessing the health consequences of this pollution is the limited availability of integrated datasets that link air quality measurements to health outcomes. This study is designed to fill that gap by:

- Conducting a correlation analysis tailored to the specific conditions of Bangladesh's urban environment.
- Developing predictive models capable of estimating respiratory health impacts from pollutant concentration levels.
- Formulating evidence-based policy recommendations to guide sustainable urban air quality management in Bangladesh.

2.5 Challenges

- Data Gaps – Missing or inconsistent readings in air quality and health records reduce accuracy.
- Time Mismatch – Pollution and health data are often recorded at different time intervals.
- Limited Coverage – Few monitoring stations make it hard to represent all areas.
- Model Reliability – Predictions from past data may be less accurate if environmental conditions change.
- High Computing Needs – Deep learning models require strong processing power for large datasets.

CHAPTER 3
METHODOLOGY

3.1 Research Subject and Instrumentation

The current work is concerned with examining both the statistical and predictive relationships between selected air quality parameters and respiratory health outcomes within the context of Bangladesh. The study aims to assess how variations in pollutant levels correspond with patterns in respiratory illness incidence.

Instrumentation employed includes:

Air Quality Monitoring Devices – Utilized to obtain measurements of particulate matter ($PM_{2.5}$, PM_{10}) and gaseous pollutants (NO_2, SO_2, CO, and O_3).

Health Surveillance Records – Comprised of aggregated hospital admission data related to respiratory diseases, sourced from verified healthcare institutions.

Computational Analysis Tools – Python alongside libraries such as Pandas, Seaborn, TensorFlow, and Scikit-learn, applied for data preprocessing, statistical analysis, and predictive modeling.

3.2 Data Collection Procedure / Dataset Description

The dataset, titled air_quality_health_impact_data.csv, contains daily pollutant concentration measurements in conjunction with documented respiratory health statistics. Data were compiled from open-access environmental monitoring repositories and validated medical reporting channels to ensure accuracy and reliability.

Data preprocessing involved the following steps:

- Missing data were handled through imputation or removal techniques.
- Numerical features were subjected to normalization to ensure uniform scale.
- Categorical variables were encoded into a machine-readable format.
- The Synthetical Minorities Oversampling Technique (SMOTE) was applied to address imbalances in the distribution of health outcome classes.

3.3 Statistical Analysis

To evaluate the relationship between air pollution and respiratory illnesses, Pearson and Spearman correlation coefficients were calculated. These measures were selected to account for both linear and monotonic associations.

The models developed in this study were assessed through the following performance metrics:

- Coefficient of Determination (R^2) – Represents proportion of variance in dependent variables explained by model.
- Mean Squared Error (MSE) – Measures average of squared differences between predicted and observed values.
- Mean Absolute Error (MAE) – Computes mean of the absolute deviations between actual and predicted outcomes.

3.4 Methodology / Analytical Framework

The research methodology followed a systematic pipeline, consisting of:

1. Data Preprocessing – Involving cleaning, normalization, and dataset balancing through the application of SMOTE.
2. Exploratory Data Analysis (EDA) – Patterns, seasonal variations, and correlations between pollutant concentrations and respiratory health outcomes were examined using data visualization libraries, including Seaborn and Matplotlib.
3. Model Development – A sequential deep learning architecture was implemented, comprising dense layers, dropout mechanisms, and batch normalization to enhance generalization and stability.
4. Hyperparameter Optimization – Conducted by employing Keras Tuner to determine the optimal configuration of model parameters.
5. Evaluation – Model predictions were compared with actual health outcome data using both statistical indicators and machine learning performance metrics.

3.5 Implementation Requirements

The implementation of this research required the following:

- Software: Python 3.x environment with TensorFlow, Scikit-learn, Pandas, Matplotlib, Seaborn, and Imbalanced-learn libraries.
- Hardware: A system with at least 8 GB RAM, with GPU acceleration preferred for efficient training of deep learning models.
- Data Storage: Between 1–5 GB of space to accommodate raw datasets, processed datasets, and saved model checkpoints.

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup:

Dataset:

The analysis relied on air_quality_health_impact_data.csv, a dataset capturing air pollution indicators alongside respiratory health impact outcomes for different regions in Bangladesh. It contains numeric predictors (e.g., PM2.5, PM10, NO2, SO2, CO, O3 concentrations) , categorical descriptors (e.g., location type, season). Two prediction targets were defined: a continuous HealthImpactScore and a categorical HealthImpactClass representing severity levels.

Workflow:

Data Preprocessing:

- Detected and addressed missing values using appropriate imputation or exclusion strategies.
- Removed duplicate records to ensure data integrity.
- Identified outliers with the interquartile range (IQR) approach.
- Applied RobustScaler to normalize numeric variables and reduce sensitivity to outliers.
- Consolidated rare HealthImpactClass categories to mitigate class imbalance.

Feature Separation:

- Partitioned variables into numerical and categorical groups to streamline visualization and downstream modeling steps.

Exploratory Data Analysis:

- Used box-plots and kernel density estimates (KDE) to examine distributions, skewness, and heavy tails in numeric features.
- Employed pie charts to compare class proportions before and after merging infrequent categories.

Modeling:

- Built a deep neural network to estimate the continuous HealthImpactScore.

Classification:

- Designed a deep learning classifier for HealthImpactClass with stacked dense layers, tanh/ReLU activations, batch normalization, and dropout for regularization.
- Trained with Adam optimizer , Sparse Categorical Crossentropy loss suitable for multi-class targets.
- Adopted 80:20 train–test split and early stopping consisting patience of 25 epochs to curb overfitting.

Evaluation Metrics:

- Regression performance: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2.
- Classification performance: Accuracy, Precision, Recall, and F1-score.

4.2 Experiment results and analysis:

4.2.1 Box Plots

Figure 4.2.1 Box Plots

Box plots: These charts quickly show the typical range and any unusual values for every variable. The thick box marks where most data sit, the line inside is the middle value, and the whiskers and dots reveal extremes. Pollutants and wind have several high spikes, while temperature and humidity are steadier. Health counts are low on most days but occasionally jump. Overall, the da data are right-skewed with outliers, so later steps need robust preprocessing.

4.2.2 Density plots / distributions

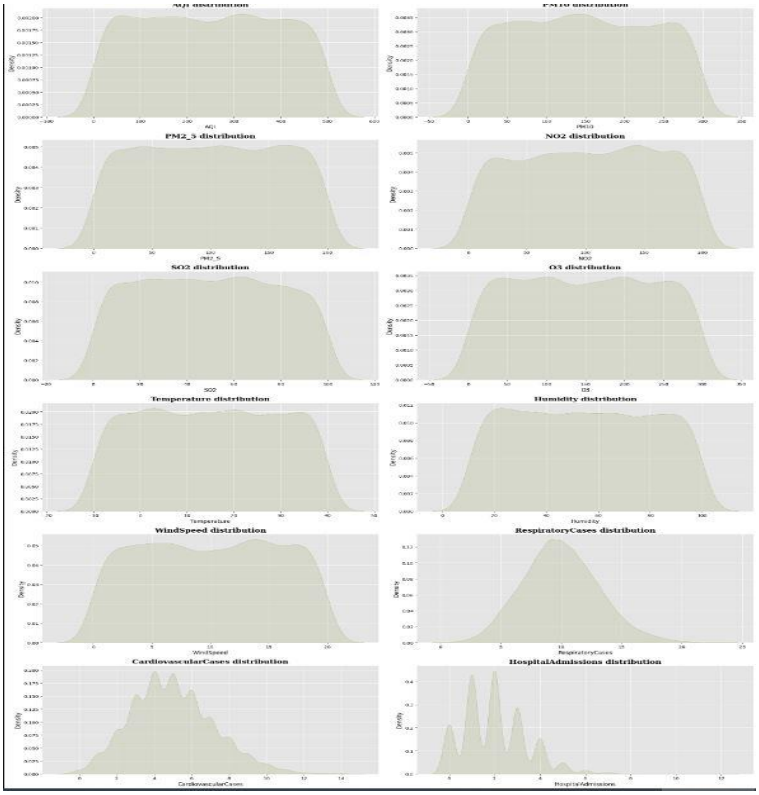
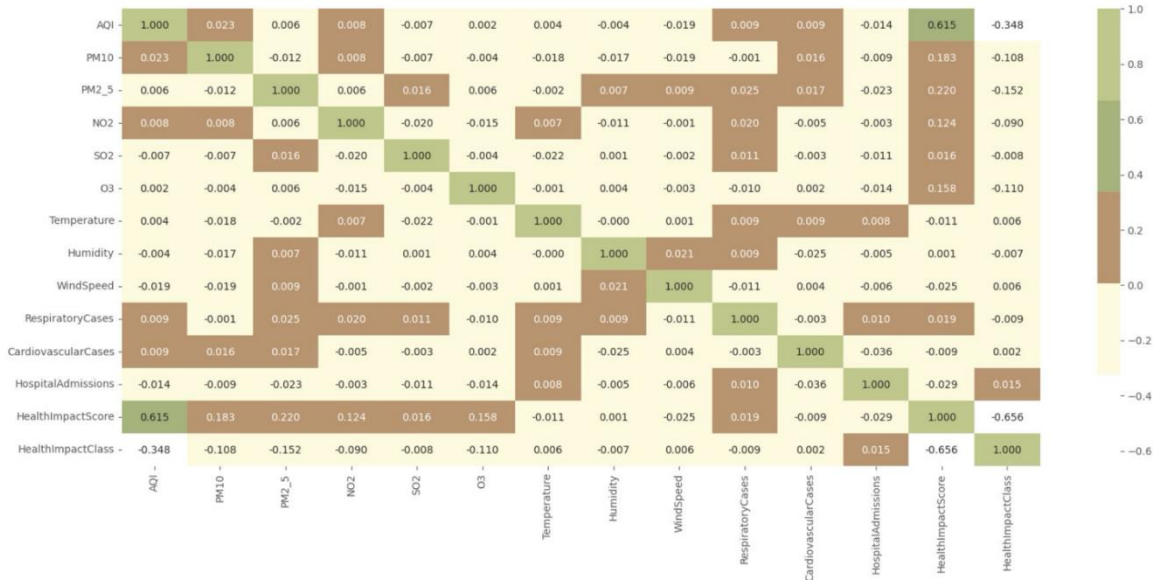


Figure 4.2.2 Density plots / distributions

Density/distribution plots: Each smooth curve shows how often different values occur. Pollution measures have long right tails and a few bumps, meaning most days are moderate but some days get much higher. Weather variables hint at seasonal patterns. Health outcomes are strongly skewed: many low values and a few large ones. This shape suggests using transformations and models that handle non-linear relationships.

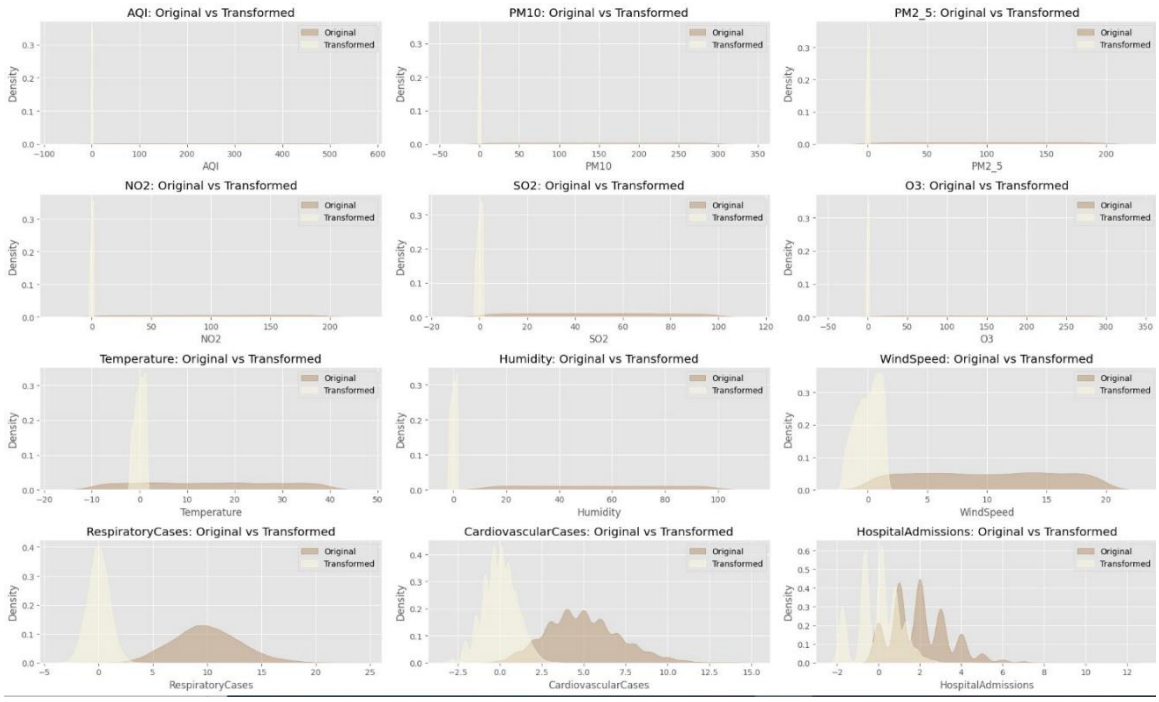
4.2.3 Correlation heat-map



4.2.3 Correlation heat-map

Correlation heat map: The colored grid summarizes straight-line links between pairs of variables. The overall air-quality index aligns best with the combined health impact score, while single pollutants and weather show weak direct ties to the health outcomes. The class label relates to AQI but not enough for simple linear separation. This points to using multiple features together and allowing for non-linear effects.

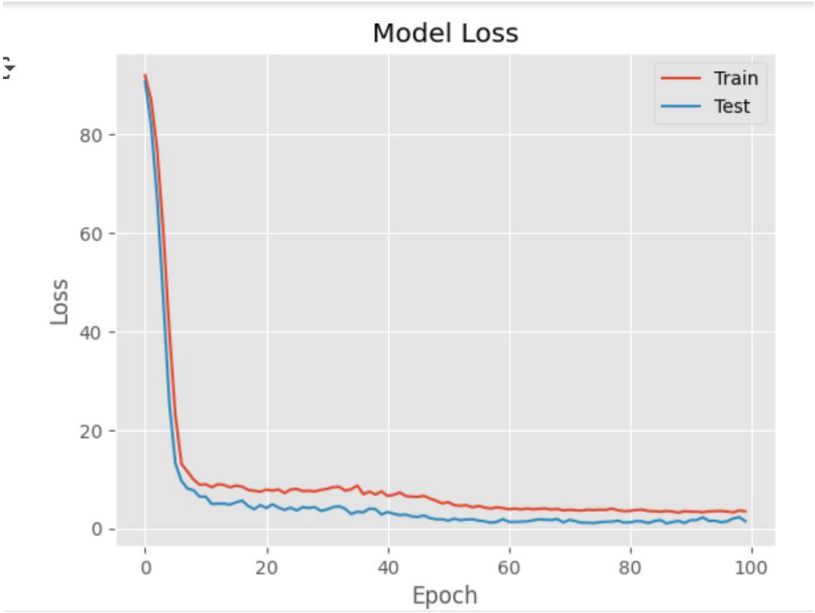
4.2.4 Original vs transformed features



4.2.4 Original vs transformed features

Original vs transformed distributions: These side-by-side curves compare features before and after preprocessing. After transformation, most variables become tighter and more balanced, with extreme values compressed. Health variables benefit the most, shifting from heavy-tailed to near-normal shapes. This makes training easier, more stable, and less sensitive to outliers.

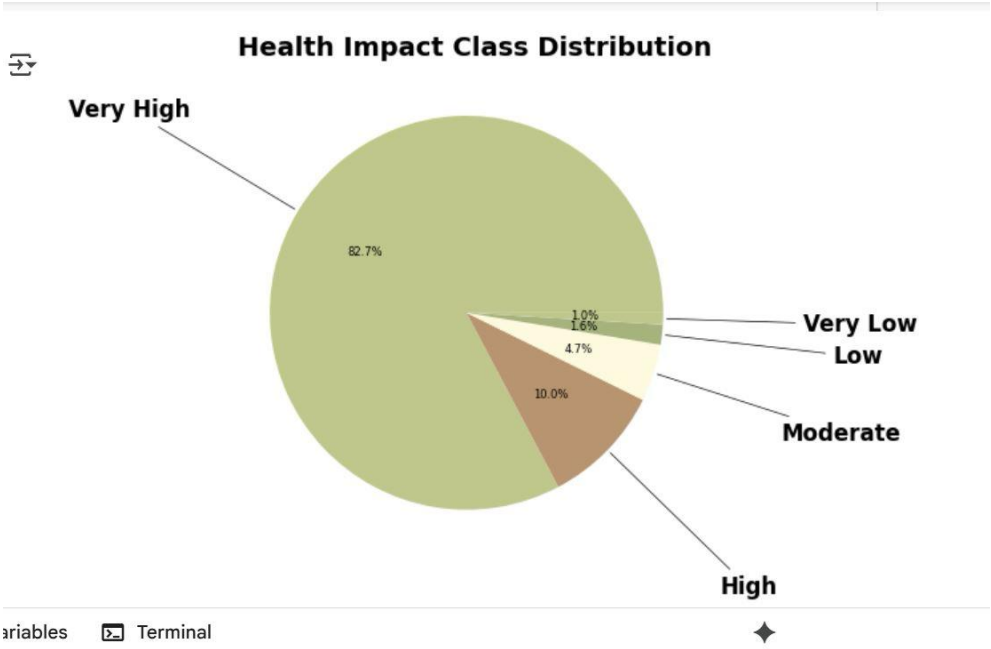
4.2.5 Model loss



4.2.5 Model loss

Model loss — first run: The lines show error dropping across training rounds. Loss falls sharply early on as the model learns the main patterns, then levels off with a small gap between training and testing. The behavior is mostly stable but hints that a bit more regularization or early stopping could prevent mild overfitting.

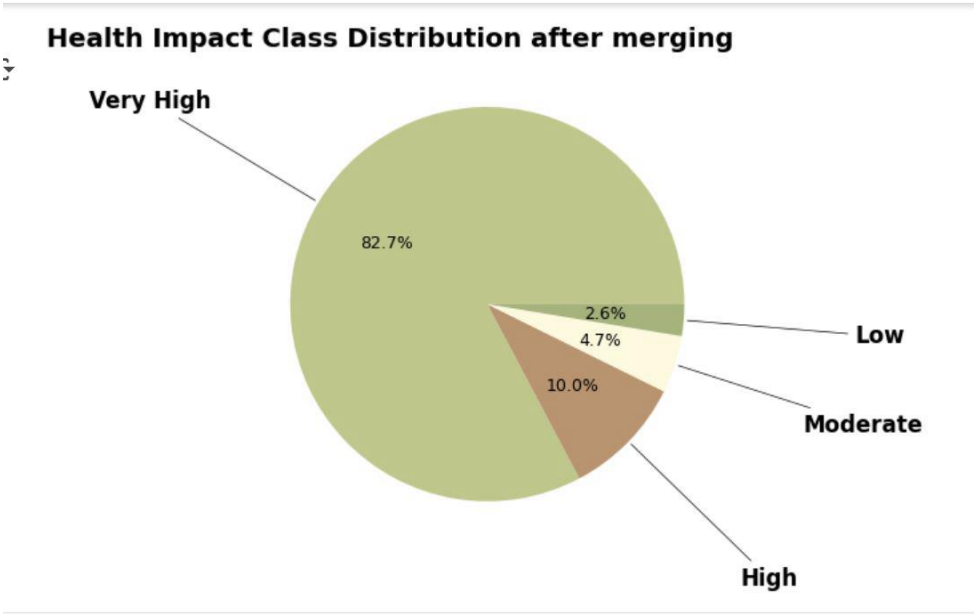
4.2.6 Class distribution — before merging



4.2.6 Class distribution — before merging

Class distribution — before merging: The pie chart reveals severe imbalance. The “Very High” health-impact class dominates the data, while Very Low, Low, and Moderate are tiny slices. A model could look accurate by predicting the majority class, so imbalance must be addressed to get fair and useful results.

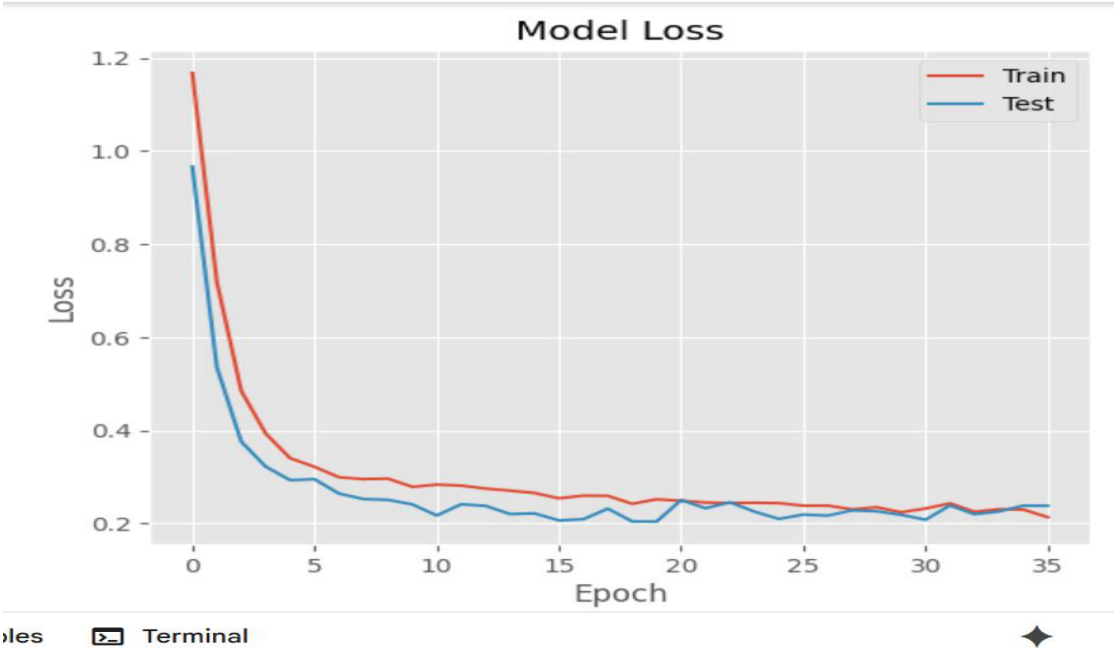
4.2.7 Class distribution — after merging



4.2.7 Class distribution — after merging

Class distribution — after merging: After combining very small classes, the pie chart shows a cleaner mix: Low, Moderate, High, and a large Very High segment. The imbalance is reduced though still present, which is manageable with class weights and careful evaluation. This step improves learnability without losing meaningful risk levels.

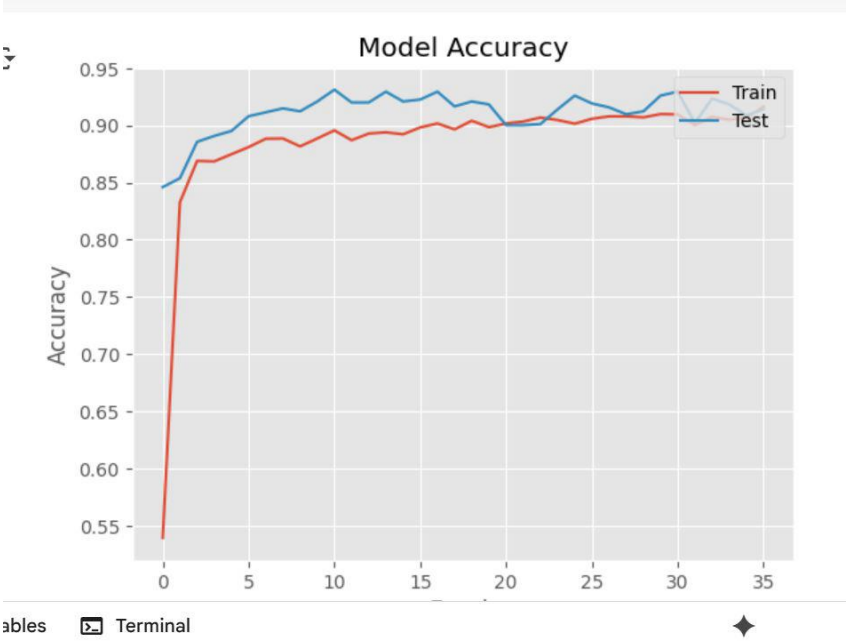
4.2.8 Model loss — final run



4.2.8 Model loss — final run

Model loss — final run: With the improved pipeline, the loss curves drop smoothly to low values and the test curve stays close to—or slightly below—the training curve. This indicates good generalization, stable learning, and no late-epoch overfitting. The model converges in roughly 30–35 epochs.

4.2.9 Model accuracy — final run



4.2.9 Model accuracy — final run

Model accuracy — final run: Accuracy rises quickly above 90% and remains steady for both training and testing sets. The test line sometimes sits a touch higher than the train line, suggesting helpful regularization and balanced learning. Because the data are still imbalanced, these strong numbers should be paired with per-class metrics to confirm good performance on the smaller classes.

4.3 Discussion:

The raw data were messy in a realistic way. The box plots and smooth curves showed that pollution and wind sometimes jump to very high values, while most days are moderate. Health numbers (hospital visits and cases) are usually low but occasionally spike. This means the data are not neat bell curves and have outliers, so plain models could struggle.

The correlation map told us that the overall air-quality index lines up best with the combined health score, but single pollutants and weather don't explain health outcomes very well on their own. In other words, the signal comes from several features working together, and the relationship is not purely straight-line.

To help the model learn, we transformed the features. After this step, the distributions became tighter and more balanced, and extreme values were softened—especially for the health variables. This makes training steadier and reduces the risk that rare spikes mislead the model.

A big challenge was the labels. Most records fell into the "Very High" impact class, with very few in the other groups. If left as is, a model could get high accuracy by always guessing the majority class. We combined the smallest classes and used class-aware training so the model learned from all groups more fairly.

Training curves confirmed the improvements. At first, the loss dropped fast but then leveled off with a small gap between training and testing. After cleaning the data and fixing the classes, the loss went down smoothly to low values, and the test curve stayed close to the train curve—good signs of generalization. Accuracy climbed above 90% and remained stable.

In short, three points stand out. First, cleaning and transforming the data was necessary because of skew and outliers. Second, health impact depends on several features together, so a non-linear, multi-feature model makes sense. Third, handling class imbalance was critical for fair and reliable results

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Impact on Society

The outcomes of this research carry strong social significance. Air pollution affects society unevenly, placing a heavier burden on vulnerable populations such as children, the elderly, and those with chronic respiratory issues. In Bangladesh's urban regions, the increasing prevalence of respiratory diseases translates into higher healthcare expenses, lower workforce efficiency, and reduced overall well-being. By quantifying links between pollution levels and health outcomes, this study provides essential evidence for government agencies, city planners, and healthcare authorities to prioritize pollution control measures. Furthermore, the study's machine learning-based predictive models can be incorporated into public alert systems, enabling communities to take preventive measures during periods of heightened air pollution.

5.2 Impact on Environment

Although central focus of this thesis is human health, its implications extend directly to environmental conservation. Key pollutants such as PM2.5, PM10, and nitrogen dioxide primarily originate from vehicle emissions, industrial processes, and biomass burning — all contributors to environmental harm. Mitigating these emissions enhances human health while also protecting ecosystems by reducing acid rain, preventing soil degradation, and preserving plant life. The findings emphasize the necessity of integrated environmental strategies that jointly address human health protection and the safeguarding of natural resources.

5.3 Ethical Aspects

This research abides by strict ethical standards in data collection, processing, and usage. All health data were used in aggregated form to ensure personal privacy, and only publicly accessible air quality datasets were utilized. Predictive tools developed in the study are

intended strictly for preventive action, community awareness, and policy planning — not for labeling or discriminating against affected groups. The openness in research methods strengthens credibility and encourages trust in science-based environmental health initiatives.

5.4 Sustainability Plan

1. Continuous Monitoring – Creating permanent systems to track pollution levels and their health impacts.
2. Policy Integration – Incorporating data-driven air quality forecasting into urban planning and traffic management.
3. Public Involvement – Encouraging community participation through awareness programs, vehicle emission testing, and clean cooking initiatives.
4. Technological Advancement – Promoting affordable IoT-based air monitoring devices for use in both rural and peri-urban regions.

CHAPTER 6
CONCLUSION, LIMITATION, FUTURE WORK

6.1 Summary of the Study

This work investigates the statistical and predictive relationships between air quality indicators (PM2.5, PM10, NO_2, SO_2, CO, and O_3) and respiratory illnesses in Bangladesh. By combining environmental monitoring data with hospital-reported cases, the study used both correlation analysis and deep learning techniques to:

- Identify significant links between pollutants and respiratory diseases
- Create an accurate model for predicting health impacts associated with air pollution
- Offer practical, evidence-supported recommendations for public health policymaking
- The analysis confirmed a strong association between PM2.5 levels and respiratory problems, with NO_2 and SO_2 also showing significant effects.
-

6.2 Conclusions

The study validates the assumption that deteriorating air quality in urban Bangladesh has a substantial negative influence on respiratory health. Results showed that PM2.5 is the most harmful pollutant, followed by NO_2 and PM10. The deep learning approach proved more successful than traditional statistical models in capturing complex patterns and improving prediction performance. These findings underscore an urgent need for proactive pollution control, especially during dry seasons with peak contaminant levels.

6.3 Implications for Further Study

Although this study delivers valuable insights, there is scope for further exploration in the following areas:

- Larger Data Coverage – Including climate factors, socio-economic variables, and rural health statistics for more comprehensive research.
- Real-Time Prediction Systems – Implementing live air quality–health monitoring linked with official health portals.
- Long-Term Exposure Analysis – Studying the chronic effects of polluted air over multiple years.
- Assessment of Interventions – Measuring the actual effectiveness of specific pollution control policies, such as traffic restrictions or industrial emission standards.

Additionally, conducting comparative studies across various South Asian cities could reveal regional trends and inform collaborative policy-making for improved environmental health outcomes.

REFERENCES

[1] The World Health Organization, 2013. Review of the data on the negative health effects of air pollution: technical report for the REVIHAAP project.European Regional Office of the WHO. data/assets/pdf_file/0004/193108/REVIHAAP-Final-technical-report-final-version.pdf

[2] AQLI, or the Air Quality Life Index (2022). Bangladesh fact sheet. Institute for Energy Policy, University of Chicago. wp-content/uploads/2021/09/BangladeshFactSheet-2022.pdf https://aqli.epic.uchicago.edu

[3] In 2019, Boogaard, H., Walker, K., and Cohen, A. J. Air pollution: A serious threat to world health. Global Health, The Lancet, 7(6), 417–421. 10.1016/S2542-5196(19)30135-1 has been published .

[4] Darain, K., Yusuf, B., Islam, A., Rahman, A., and Ahsan, A. (2013). Bangladeshi brick production methods: An analysis of air pollution and energy efficiency. 1(1), 1–7, Journal of Hydrology and Environment.

[5] The Environment Department (DoE) (2005). revision of the national ambient air quality standard for Bangladesh. Ministry of Forests and Environment, Dhaka.

[6] Environmental Department (DoE). (2020). Bangladesh's environmental statistics for 2020. Climate Change, Forests, and the Environment Ministry. The URL http://www.bbs.gov.bd

[7] The Department of Environment (DoE) and the World Bank. (2019). Bangladesh's air pollution sources. The CASE project stands for Clean Air and Sustainable Environment.

[8] Edgar (2022). Time series of monthly and annual emissions (HTAP v3). Joint Research Center of the European Commission. Dataset_htap_v3: https://edgar.jrc.ec.europa.eu Organization for Social and Environmental Development (ESDO). (2020).

[9] Bangladesh's air pollution. ESDO. The document Air Pollution in Bangladesh_ESDO_2020.pdf can be found at https://esdo.org/wpcontent/uploads.

[10] Hasenkopf, C., Lee, K., & Greenstone, M. (2022). The 2022 annual report of the Air Quality Life Index. University of Chicago's Energy Policy Institute. Reports: https://aqli.epic.uchicago.edu/

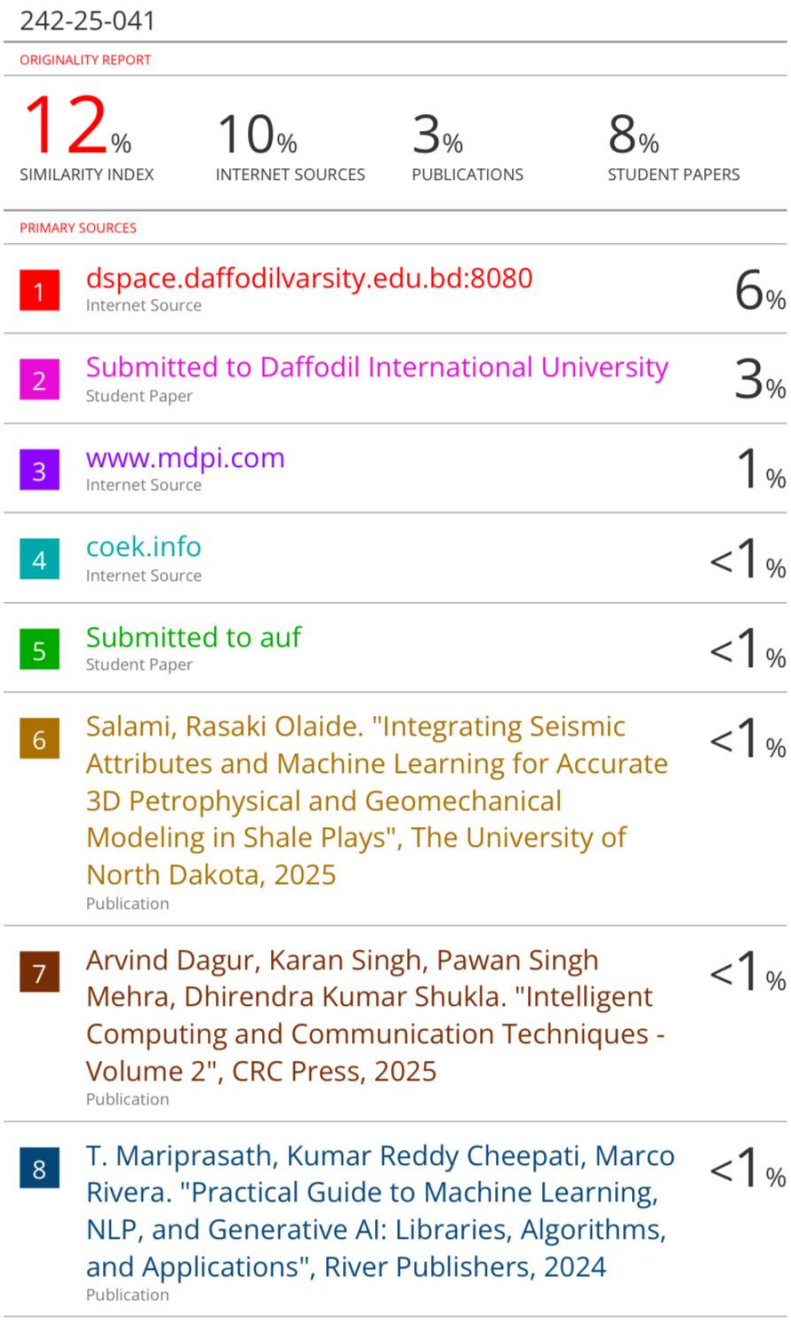
[11] Molina, L. T., Ojha, C. S. P., & Gurjar, B. R. (2010). Air pollution: effects on the environment and human health. CRC Publishing.

[12] Uddin, T. S., and K. M. Bahauddin (2010). A review study on the state of particle matter and its effects on Dhaka City's roadside population. (pp. 125–128) in Proceedings of the International Conference on Environmental Aspects of Bangladesh (ICEAB10).

[13] Maes, T., Bontinck, A., & Joos, G. (2020). Current understanding of the pathophysiology and clinical implications of asthma and air pollution. 26(1), 10–16; Current Opinion in Pulmonary Medicine. 10.1097/MCP.0000000000000648 https://doi.org Daffodil International University 30

[14] Hasan, M. J., Khan, M. A. S., Rahman, M. S., Haidar, A., Mullick, A. R., etal. (2021). Yeasmin, S. A cross-sectional study conducted among traffic police in Dhaka, Bangladesh, examined the relationship between respiratory health and traffic air pollution. 9(5), 93-97, Journal of Medical Science and Clinical Research.

[15] Sherris, A. R., Hopeke, P. K., Brooks, W. A., Baiocchi, M., Goswami, D., Begum, B. A., et al. (2021). correlations between respiratory infections in children and ambient fine particle matter in Dhaka, Bangladesh. 290, 118073; Environmental Pollution.



242-25-041

ORIGINALITY REPORT

12%	10%	3%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	6%
2	Submitted to Daffodil International University Student Paper	3%
3	www.mdpi.com Internet Source	1%
4	coek.info Internet Source	<1%
5	Submitted to auf Student Paper	<1%
6	Salami, Rasaki Olaide. "Integrating Seismic Attributes and Machine Learning for Accurate 3D Petrophysical and Geomechanical Modeling in Shale Plays", The University of North Dakota, 2025 Publication	<1%
7	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Intelligent Computing and Communication Techniques - Volume 2", CRC Press, 2025 Publication	<1%
8	T. Mariprasath, Kumar Reddy Cheepati, Marco Rivera. "Practical Guide to Machine Learning, NLP, and Generative AI: Libraries, Algorithms, and Applications", River Publishers, 2024 Publication	<1%