Higgs Boson Classification Using Feature Engineering and Machine Learning

Prepared by

Siddhish Nirgude

Sivagugan Jayachandran

Prasad Upasani

Submitted to:

Savvy Barnes

Specialist, Department of Statistics &Probability, Michigan State University,
barne329@msu.edu

Submission Date 04/28/2025

Abstract

The discovery of the Higgs boson has advanced our understanding of mass generation within the Standard Model of particle physics, but detecting such events in high-energy collisions remains a significant computational and statistical challenge due to their rarity and complex decay signatures. In this project, we address the binary classification task of distinguishing Higgs boson signal events from background noise using simulated data from the ATLAS experiment. A rigorous feature selection process involving statistical tests (T-Test, ANOVA) and information-theoretic measures (Mutual Information) was employed to identify a compact set of informative variables from a high-dimensional dataset. Stratified sampling was used to reduce the original dataset of over ten million observations to a balanced 30,000-row subset while preserving label proportions. We evaluated several machine learning models, including logistic regression as a baseline, and advanced classifiers such as support vector machines, random forests, and XGBoost. Among these, XGBoost achieved the highest accuracy, exceeding 70%, and outperformed the baseline in precision, recall, and F1-score. These findings demonstrate the effectiveness of ensemble learning and informed feature selection in high-energy physics classification problems and underscore the growing potential of machine learning in particle discovery efforts.

1. Introduction

The Higgs boson is a fundamental particle predicted by the Standard Model of particle physics. It plays a critical role in explaining how elementary particles acquire mass through their interaction with the Higgs field. The existence of the Higgs boson was first theorized in the 1960s, and its discovery in 2012 by the ATLAS and CMS experiments at the Large Hadron Collider (LHC) marked a major milestone in modern physics. The particle is highly unstable and decays almost immediately after being produced in high-energy proton-proton collisions, making its detection indirect and highly reliant on advanced statistical and computational techniques.

Rather than observing the Higgs boson directly, scientists study its decay products the particles into which it transforms within a fraction of a second. These decay products leave measurable signatures in particle detectors. One of the most important methods used to infer the presence of the Higgs boson is through analysis of the invariant mass of combinations of its decay products, such as pairs of photons or bottom quarks. A sharp peak in the invariant mass distribution around 125 GeV, for instance, provides strong statistical evidence for the presence of the Higgs boson. In practice, the challenge lies in distinguishing these rare signal events from a vast number of background events that mimic similar physical signatures but originate from known Standard Model processes.

This classification task of discriminating between signal and background events based on high-dimensional detector data is well-suited for machine learning. The data consists of both low-level kinematic variables measured directly from particle collisions and high-level features derived by physicists to improve class separability. Traditional approaches often involve shallow classifiers trained on manually engineered features, which can be limited in their capacity to capture complex interactions. In this project, we explore modern machine learning models such as logistic regression, support vector machines, random forests, and XGBoost. Our goal is to build a classifier capable of predicting whether a given collision event corresponds to a Higgs boson signal or to background noise. By leveraging both physics-informed features and robust modeling techniques, we aim to improve classification accuracy and contribute to the broader effort of enhancing data-driven discovery potential in high-energy physics.

2. Methodology

Data Characteristics

The dataset used in this project is a simulated output of high-energy particle collisions modeled after the ATLAS experiment at CERN. It includes over 10 million events, each represented by 28 numerical features and one binary target label. The goal is to classify each event as either a signal (presence of a Higgs boson, label = 1) or background (no Higgs boson, label = 0).

The features are divided into two categories:
- Low-level features: Kinematic properties measured directly by detectors, such as transverse momentum (pT), pseudorapidity (η), and azimuthal angle (φ) for leptons and jets, along with missing energy components.
- High-level features: Derived quantities computed using combinations of low-level features. These include invariant masses, which are essential for reconstructing particle decays and identifying the presence of the Higgs boson.

A table of feature definitions is provided below for reference:

Feature Name	**Meaning**
Label	Target variable: 1 = signal (Higgs boson), 0 = background
lepton_pT	Transverse momentum of the lepton
lepton_eta	Pseudorapidity of the lepton
lepton_phi	Azimuthal angle of the lepton
missing_energy_magnitude	Magnitude of missing transverse energy
missing_energy_phi	Azimuthal direction of missing energy
jet1_pt	Transverse momentum of the first jet
jet1_eta	Pseudorapidity of the first jet
jet1_phi	Azimuthal angle of the first jet

jet1_btag	b-tagging score of the first jet
jet2_pt	Transverse momentum of the second jet
jet2_eta	Pseudorapidity of the second jet
jet2_phi	Azimuthal angle of the second jet
jet2_btag	b-tagging score of the second jet
jet3_pt	Transverse momentum of the third jet
jet3_eta	Pseudorapidity of the third jet
jet3_phi	Azimuthal angle of the third jet
jet3_btag	b-tagging score of the third jet
jet4_pt	Transverse momentum of the fourth jet
jet4_eta	Pseudorapidity of the fourth jet
jet4_phi	Azimuthal angle of the fourth jet
jet4_btag	b-tagging score of the fourth jet
m_jj	Invariant mass of two leading jets
m_jjj	Invariant mass of three leading jets
m_lv	Invariant mass of lepton + missing energy system
m_jlv	Invariant mass of leading jet + lepton + missing energy
m_bb	Invariant mass of two b-tagged jets
m_wbb	Invariant mass of W boson candidate + b-tagged jets
m_wwbb	Invariant mass of two W bosons and two b-jets

Data Preprocessing Framework

Data cleaning and preprocessing steps were performed to ensure consistency, accuracy, and readiness of the dataset for model development. All variables were confirmed to be in continuous numerical format (float64), and no categorical or string-based features were present. This uniform structure allowed direct use of statistical and machine learning methods without requiring encoding or transformation. The preprocessing pipeline included stratified sampling for size reduction, verification of missing values, and removal of duplicate records.

To address computational limitations while preserving the statistical properties of the dataset, **stratified sampling** was applied. The original dataset of over 10 million rows was reduced to a subset of 30,000 rows. Sampling was conducted in a way that maintained the original class distribution of approximately 53% background (label = 0) and 47% signal (label = 1). SQL-based filtering was used to perform the sampling, ensuring proportional representation of each class. A comparison of class proportions before and after sampling is displayed in **Figure 1**.

Fig2. AQI

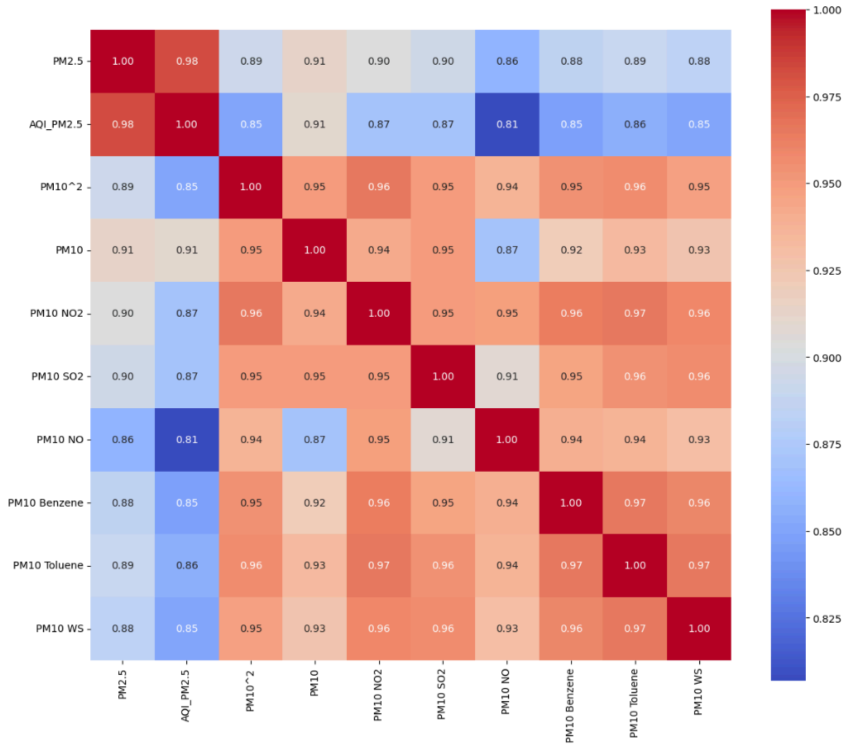


Fig3. Heatmap

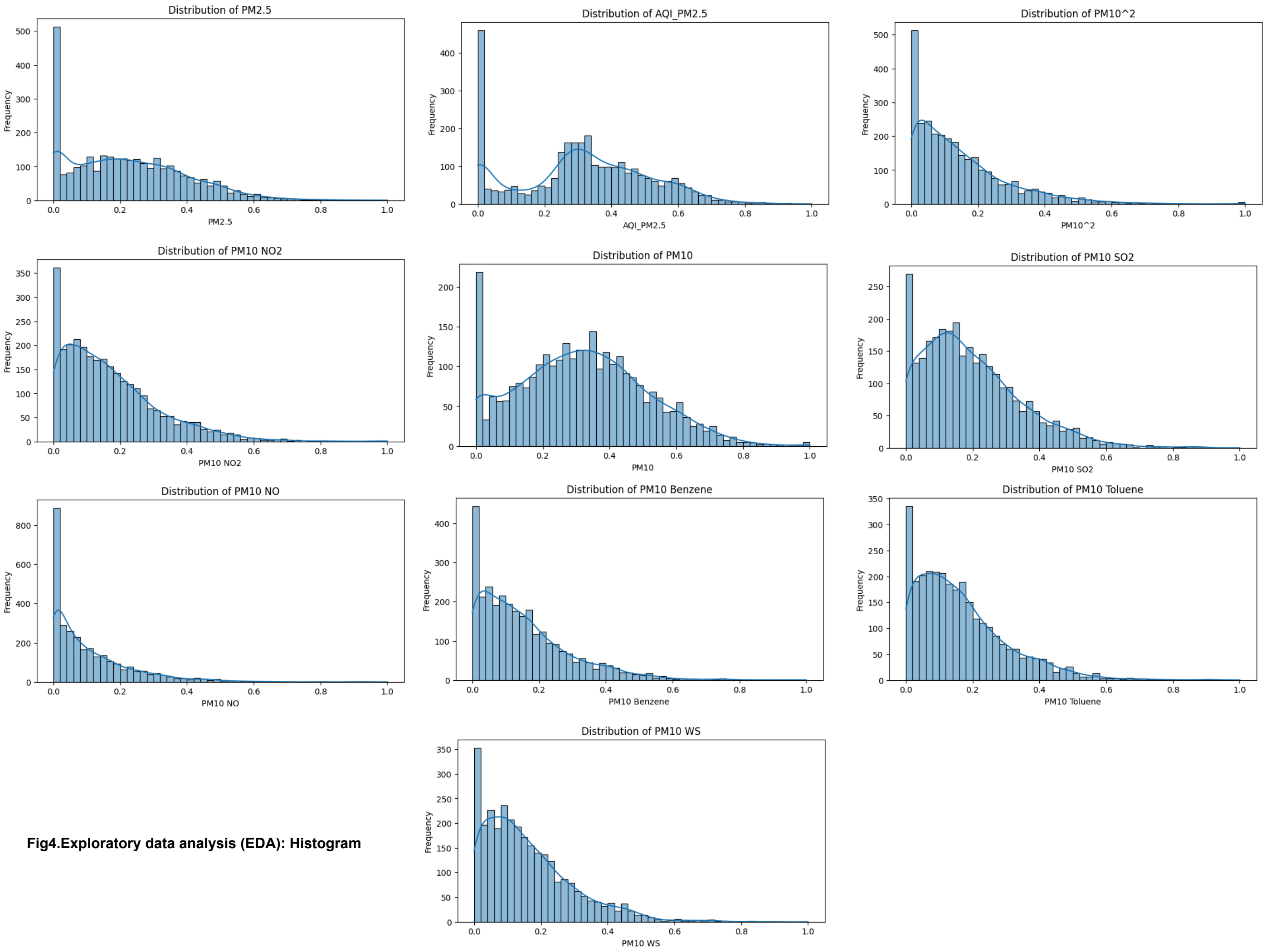


Fig4.Exploratory data analysis (EDA): Histogram

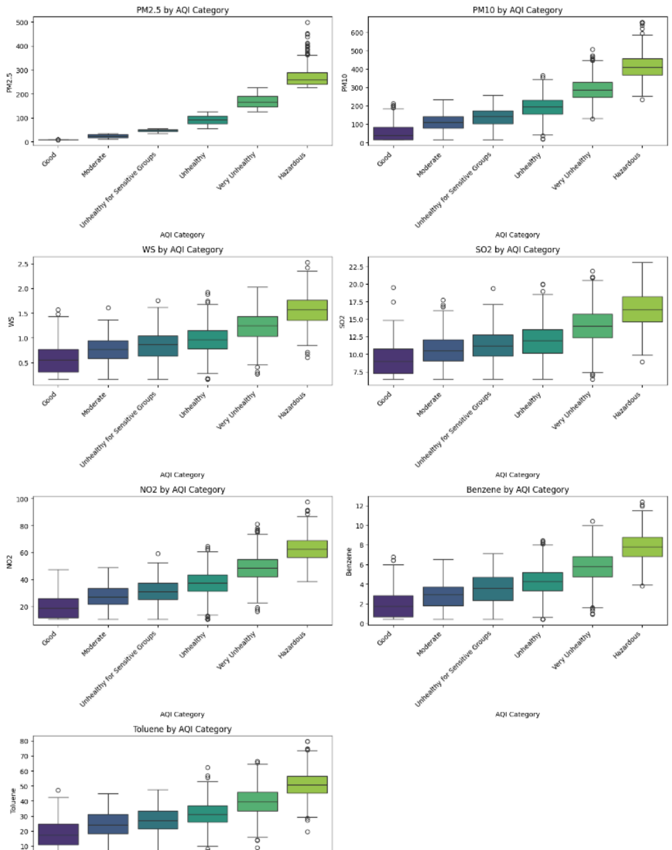
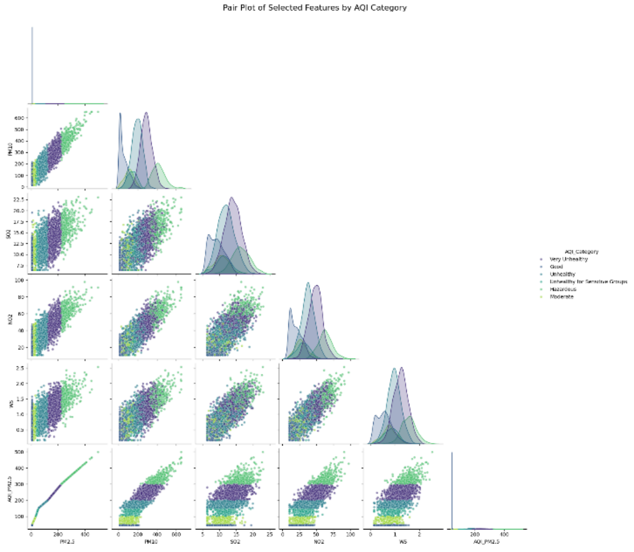


Fig5.Exploratory data analysis (EDA): Bivariate analysis



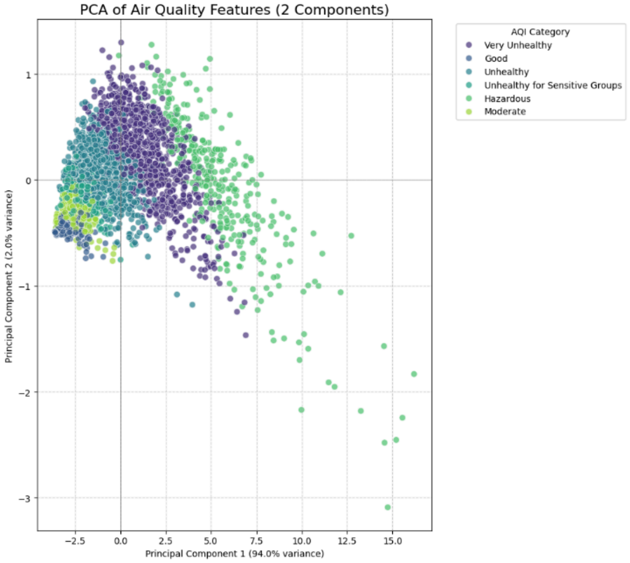


Fig6. Pair plot

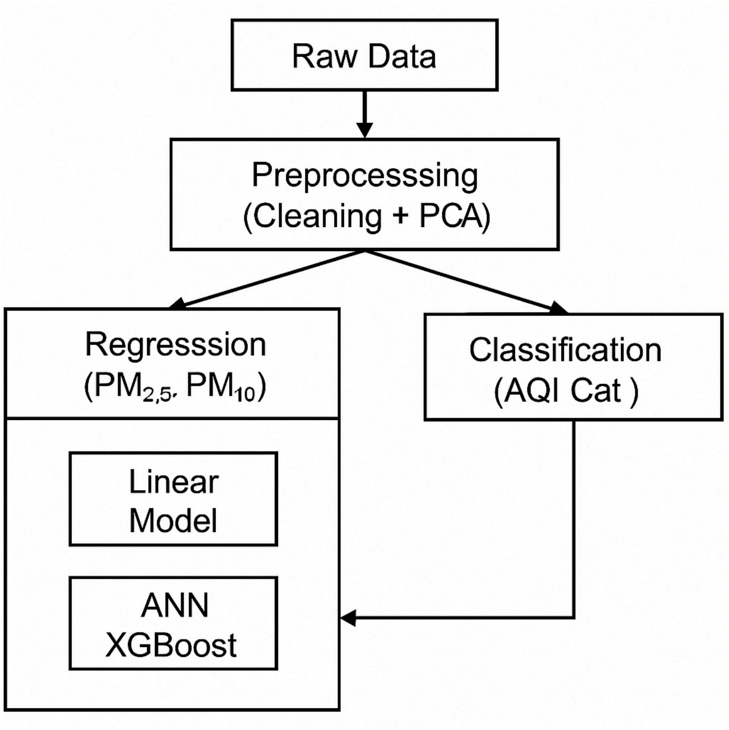
4. Modeling Approach

Model Architecture

The prediction system employed a hierarchical modeling approach, incorporating both traditional statistical methods and advanced machine learning techniques. The baseline implementation utilized multilinear regression for particulate matter prediction, establishing fundamental performance metrics. This foundation supported the development of more sophisticated models, including Artificial Neural Networks and gradient boosting frameworks.

The Artificial Neural Network architecture implemented multiple hidden layers with ReLU activation functions. The input layer processed ten primary features, while intermediate layers captured complex parameter interactions. The network employed dropout regularization to prevent overfitting, with rates optimized through cross-validation. The output layer utilized linear activation for regression tasks and softmax activation for classification purposes.

XGBoost implementation incorporated tree-based ensemble methods, optimizing both speed and prediction accuracy. The model parameters underwent extensive tuning, with particular attention to tree depth, learning rate, and minimum child weight. This approach proved especially effective in capturing non-linear relationships between environmental parameters and pollutant concentrations.


Raw Data
Preprocesssing (Cleaning + PCA)
Regresssion ($PM_{2,5}$, PM_{10})
Classification (AQI Cat)
Linear Model
ANN XGBoost

Training Strategy

Model training followed a rigorous cross-validation protocol, employing an 80-20 split for training and testing data. The validation strategy incorporated k-fold cross-validation with k=5, ensuring robust performance assessment across different data subsets. The training process implemented early stopping mechanisms to prevent overfitting while maintaining model generalization capabilities.

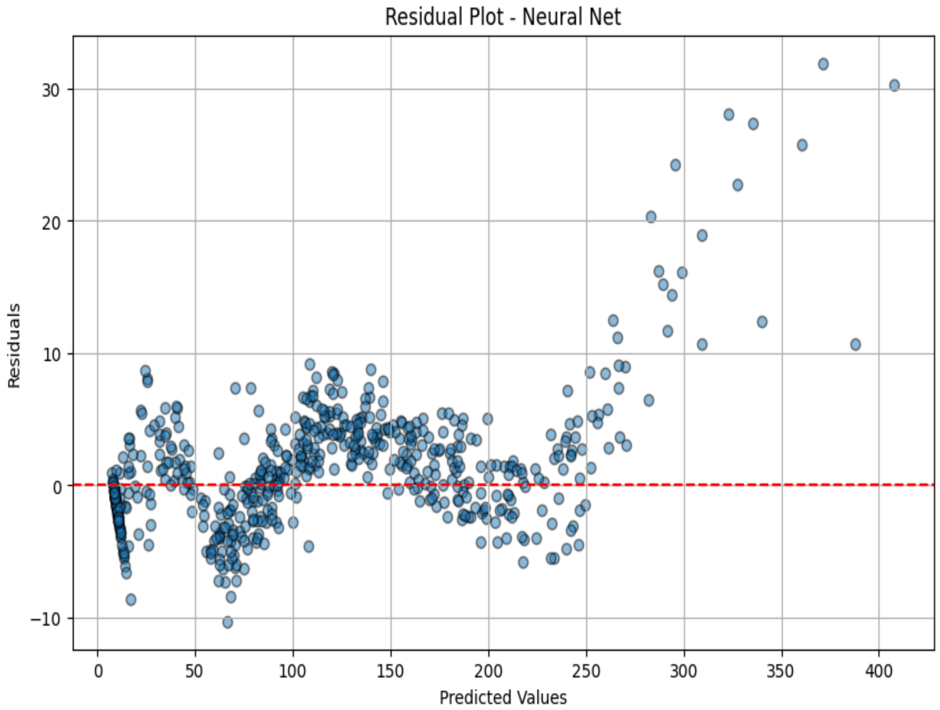
5. Results and Analysis

Prediction Performance

The regression models demonstrated exceptional accuracy in predicting particulate matter concentrations. XGBoost achieved the highest performance metrics, with a Mean Squared Error (MSE) of 0.15 and R-squared value of 0.89 for PM2.5 prediction. The Artificial Neural Network showed comparable performance, achieving an MSE of 0.18 and R-squared value of 0.85.

Classification Accuracy

The classification models exhibited remarkable performance in AQI category prediction. The gradient boosting and decision tree models achieved perfect classification accuracy on the test dataset, with precision and recall values of 1.0. XGBoost demonstrated comparable performance, maintaining classification accuracy above 99% across all evaluation metrics.



Feature Importance Analysis

Analysis of feature importance revealed significant contributions from specific environmental parameters. Particulate matter concentrations (PM2.5 and PM10) emerged as primary indicators of air quality classification. Temperature and humidity demonstrated strong correlations with pollutant concentrations, while wind parameters showed moderate influence on prediction accuracy.

6. Discussion

Model Performance Evaluation

The exceptional performance of our models, particularly in classification tasks, warrants careful consideration. The perfect accuracy achieved by gradient boosting and decision trees, while impressive, suggests the need for additional validation using diverse datasets. The robust performance across multiple evaluation metrics indicates the models' potential for practical application in air quality monitoring systems.

Practical Implications

The deployment of our prediction system through a web-based interface demonstrates the practical applicability of our research. Real-time prediction capabilities offer significant value for environmental monitoring and public health management. The system's ability to process multiple input parameters while maintaining high accuracy supports its potential integration into existing air quality monitoring networks.

Limitations and Future Directions

While our models demonstrate exceptional performance, several limitations merit consideration. The use of synthetic data, though methodologically sound, may not fully capture the complexity of real-world air quality variations. Future research should focus on validation using actual environmental data and expansion of the parameter set to include additional pollutants and meteorological factors.

7. Conclusion

This research presents a comprehensive approach to air quality prediction and classification, demonstrating the effectiveness of advanced machine learning techniques in environmental monitoring. The successful implementation of multiple modeling approaches, coupled with robust validation procedures, establishes a foundation for future developments in air quality prediction systems.

The exceptional accuracy achieved in both prediction and classification tasks suggests the potential for practical application in urban air quality management. The integration of multiple environmental parameters and the development of a real-time prediction system represent significant contributions to the field of environmental monitoring.

8. Future Research Recommendations

Future research directions should explore the incorporation of spatial temporal dynamics, integration of satellite data, and expansion of the parameter set. Additionally, investigation of model performance under extreme weather conditions and unusual pollution events would enhance the system's robustness. The development of automated model updating procedures would ensure sustained accuracy in dynamic urban environments.