Improving Student Experience with a Flexible Machine Learning System for Tracking Air Quality

by

Priota Das
22101610
Md. Towfiqul Alam Khan
20301226
Nabiha Binte Nasim
22101727
Shoumik Das
21101046
Sultana Haque
20301396

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
October, 2025

Declaration

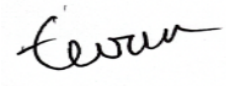
It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

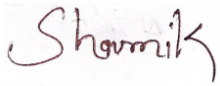
Student's Full Name & Signature:

Priota Das
22101610

Md. Towfiqul Alam Khan
20301226



Nabiha Binte Nasim
22101727



Shoumik Das
21101046

Sultana Haque
20301396

Approval

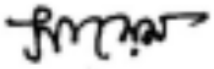
The thesis titled "Improving Student Experience with a Flexible Machine Learning System for Tracking Air Quality" submitted by

1. Priota Das (22101610)
2. Md. Towfiqul Alam Khan (20301226)
3. Nabiha Binte Nasim (22101727)
4. Shoumik Das (21101046)
5. Sultana Haque (20301396)

Of Summer, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on October 10, 2025.

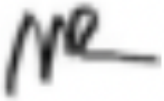
Examining Committee:

Supervisor:
(Member)



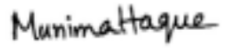
Dr Abu Mohammad Khan
Associate Professor
Department of Computer Science and Engineering
Brac University

Secondary Supervisor:
(member)



Annajiat Alim Rasel
Senior Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(member)



Dr Munima Haque
Associate Professor, Biotechnology Program
Department of Mathematics and Natural Sciences (MNS)
Brac University

Thesis Coordinator:
(member)

Dr Md Golam Rabiul Alam
Thesis Coordinator
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Air pollution is significantly a significant silent warning for human life in this digitized era. This pollution is a horrible threat for public health. Students spend most of the day at their educational premises, so the air quality of any educational setting has a wide impact on students' health, which is a great concern.

This research focuses on building a hybrid model combining machine learning and deep learning techniques to predict Air Quality Index (AQI) levels more accurately than existing models. The dataset consists of 1632 data points with key pollutants Carbon Dioxide (CO_2), Carbon Monoxide (CO), Total Volatile Organic Compounds (TVOC), Formaldehyde (HCHO), Particulate matter 0.3 ($PM_{0.3}$), Ultrafine particulate matter($PM_{1.0}$), Fine particulate matter ($PM_{2.5}$), and Particulate Matter ($PM_{10.0}$) collected from 18 different locations on campus over 7 months (January 2025 to July 2025).Multiple machine learning (ML) models, e.g., Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Categorical Boosting (CatBoost), Light Gradient Boosting Machine (LightGBM), and Support Vector Machine (SVM), as well as deep learning (DL) models, including Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), and a proposed hybrid model, were trained and evaluated using standard metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC curve after appropriate data preprocessing. Additionally, as a part of explainability integration, SHapley Additive explanations (SHAP) were observed for the proposed model in this study. The forced Shap plot for different samples in each of the five classes of AQI classification both correctly and incorrectly predicted the class to detect the feature contribution behind the model's prediction.

Experimental results show that the proposed hybrid model accuracy is highest among all used machine learning and deep learning models. CatBoost, LightGBM, and Random Forest are the top three models in machine learning, and SVM is the highest in deep learning models.

The proposed model air quality classification system effectively combines machine learning with explainable AI methods to improve experience and build trust in environmental monitoring. It shows promising potential for real-time use in educational settings to increase awareness and enable informed health decisions. However, this study has few limitations, e.g.an imbalanceded dataset, lack of data in each class, and model's limitations, e.g., overfitting risk, lack of tuning, and so on; future efforts will concentrate on real-world validation and refining the model.

Dedication

It is our genuine gratefulness and warmest regard that we dedicate this work to all our loved ones for their love and inspiration.

Keywords

Air Quality, IAQ (indoor air quality), ML (Machine learning) framework, SHAP, Deep learning, Hybrid Methods, Anomaly Detection, Seasonal Trends, EDA.

Acknowledgement

We express our gratitude to all who uphold us with their proper guidance throughout the completion of our thesis journey. We are grateful to our thesis supervisors, Dr. Abu Mohammad Khan, Annajiat Alim Rasel, and co-supervisor Dr. Munima Haque, especially for their patience and dedication in helping us in our thesis to develop our analytical skills. Their constant guidance and sagacious feedback will be a blessing for our thesis. Additionally, we would like to thank our respectful faculty members who helped us learn the knowledge from the root to the roof.

As a part of international collaboration, this work was partially funded by Brac University Research Seed Giant Initiative (RSGI) 2024-2025 under a research project entitled "A Multi-level and Multi-seasonal Study of Air Quality in the University and Surrounding Areas". We gratefully appreciate this support that helped us carry on and expand our research.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Context

Air pollution holds the title of being one of the most critical environmental and public health problems of the 21st century. Long-term exposure to poor-quality air has been causing respiratory illness, allergies, fatigue, and multiple health issues for years. For students, this can lead to reduced concentration, lower academic performance, and overall a threat to their well-being due to spending hours inside campus premises with indoor pollutants such as carbon monoxide(CO), particulate matter(PM) etc.
On most university campuses, there is little awareness about the poor air quality inside the campus premises such as classrooms, lab rooms, cafeteria or common rooms. The traditional air quality monitoring systems usually cannot detect local variations within the campus because most of these are limited to providing general city level air quality data. This calls for an intelligent and real time air quality monitoring system that can read real time data from different location inside the campus and give students an idea of what they are breathing.

1.2 Problem Statement

The usual AQI systems are not designed for personal use or campus based use. They often show delayed or generalized results not providing the proper explanation of how certain air quality levels are assigned. As a result, the user is unsure of what the data means or how to respond to such results. There is also the lack of real time feedback and explainability in these systems which limits practical use of it, especially in student oriented environments.

1.3 Objectives

The main goal of this study is to create and design an explainable machine learning model that can classify air quality levels accurately. The specific objectives to create and compare different machine learning models for classifying AQI levels based on real-time sensor data, help users to make better daily activity and health decisions

using understandable and transparent insights from the model.

1.4 Contribution

This study has several valuable contributions to drawing inferences for air quality monitoring and intelligent environmental analysis.

- Creation of a new dataset:

  Each of the air quality data used for this research was directly measured in real-time by the authors from multiple indoor and outdoor locations on Brac University campus. This self-developed dataset has both temporal and spatial data variation. This offers a unique foundation for future studies of real-world air quality analysis.

- Building Hybrid model:

  We developed and implemented a hybrid model that not only predicts AQI values but also estimates and fills in missing values. This is something that conventional devices cannot do. Standard sensors rely on real time readings only but our model integrates both deep learning and machine learning to fill in missing data and ensure consistent AQI prediction.

- Model evaluation:

  We compared performance of our hybrid model with various machine learning and deep learning methods with key metrics such as precision, recall, F1-score, ROC-AUC curve, and confusion matrix. This comparison highlighted the outstanding accuracy and stability of our model.

- Explainability and interpretability:

  We employed SHapley Additive exPlanations (SHAP) to illustrate the contribution of each pollutant feature to AQI classification. This explainability feature helped facilitate a better understanding of how the model projected and confirmed its interpretive accuracy.

1.5 Scope and Limitations

The research focuses on indoor air quality in the BRAC University campus, which is an urban campus. Data were collected with MT15-Tuya multifunctional air quality sensors at different indoor and outdoor locations. The main pollutants being monitored are $PM_{2.5}$, PM_{10}, CO_2, CO, TVOC, and HCHO, along with temperature and humidity.
However, pollutants like O_3, SO_2, and NO_2 were excluded because sensors are unable to measure them. The task is also focused on model explainability and training rather than mass deployment. Therefore, while the system provides valuable insights, its outputs are specific to this dataset and may fail to represent other regions or environments.

1.6 Significance of the Study

The study contributes to building an intelligent campus systems by integrating machine learning, sensor technology, and explainable AI. The system, through providing understandable information on air quality levels, enables students, scholars, and university administrators to make educated health and behavioral decisions. Furthermore, the application of SHAP and LIME provides transparency and accountability in predictive models and enables the construction of user trust in AI-based environmental surveillance. Outside of the academic environment, the system can become a modular platform for smart air quality systems in schools or in urban spaces.

1.7 Thesis Organization

This thesis is divided into seven chapters, each depending on the other one to provide a complete understanding of the analysis, development, and evaluation of an explainable machine learning system for campus air quality monitoring.

- Chapter 1 – Introduction
  This chapter establishes the background of the research, problem statement and objectives. It also states the successes, limitations, and significance of the study.
- Chapter 2 – Literature Review
  This chapter reviews relevant papers on air quality monitoring standards, systems, and techniques. It discusses existing AQI models that are used in different papers with a similar goal, compares different machine learning approaches used in air quality prediction, highlights AI techniques and identify the research gaps that motivates this study.
- Chapter 3 – Methodology
  This chapter describes the workflow of the system and data collection process overall. It introduces the dataset and the features, types of pollutants, measuring methods, and preprocessing processes. This chapter also describes the machine learning models, performance metrics, as well as implementation tools and frameworks.
- Chapter 4 – Results and Analysis
  This chapter presents the result of all models used based on different metrics. It compares classification results of all algorithms and visualizes them with confusion matrices and performance charts.
- Chapter 5 – Discussion
  In this chapter the results explain why certain models performed better and how feature importance aligns with environmental factors. This chapter also includes the practical implications of the findings.

- Chapter 6 – Limitations
  Here the limitations faced during the study are discussed. This includes dataset limitations, model-specific issues and explainability challenges in machine learning models. It also discusses system-level, ethical, and generalization concerns.

- Chapter 7 – Conclusion and Future Work
  This chapter gives a summary of the main findings and contributions of the study. This chapter also adds the performance of the proposed machine learning and explainability method. It also focuses on the study's limitations and directions for the future.

Chapter 2

Literature Review

2.1 AQI Systems and Standards

In the paper 'Air Quality Index (AQI) variations and spatial distribution in Bangladesh from 2014 to 2019', data from different Continuous Air Monitoring Stations were analysed by Majumder et al. to calculate geographical and seasonal pollution trends [1]. Findings were that Dhaka, Narayanganj, and Gazipur had extremely unhealthy air quality in winter and pre-monsoon periods, with Sylhet and Barisal comparatively better. Authors expressed a strong positive correlation between $PM_{2.5}$ levels and AQI with increased levels of pollution during dry months and relatively lesser levels during the monsoon period due to rain washout effects. The findings support the intensity of air pollution in cities, indicate the role of industrialization and traffic emissions in lowering the quality of air, and confirm the use of AQI as a valid tool for measuring environmental health risks and guiding public health policies.

Majumder et al. (2023) detailed the seasonal and spatial changes in AQI in six districts of Bangladesh for 2014-2019, and a commendable effort was made in diffusing the AQI dataset, multiple gaps remain in their work [1]. First, the work only discusses the correlation of AQI and $PM_{2.5}$, and does not take into consideration other potential harmful pollutants such as NO_2 and O_3. Second, the work only examines and analyzes AQI data with a temporal resolution of a month and in a seasonal manner. The examination of daily and diurnal patterns in AQI variance are critical for understanding the exposure in the fine scale. Third, the work, despite studying six districts, spatially misses large areas of rural and rural-urban fringe communities, which might possess differing pollutant dynamics. Fourth, beyond precipitation, other meteorological elements like wind speed and humidity, and the height of the atmospheric boundary layer, are critical elements in the dispersion of pollutants and variation of AQI. These elements need to be included in the analysis. Lastly, the work does not incorporate health outcome data, which weakens its policy recommendation.

2.2 Air Quality Monitoring Techniques (Satellite / Sensor / Hybrid)

In 'Satellite Monitoring of Air Quality and Health', it was explored by Holloway et al. (2021) how remote sensing data like AOD and tropospheric NO_2 can serve as reliable proxies for pollutant concentrations in regions lacking dense ground networks [2]. Satellite-based monitoring systems are important for air quality and health research to monitor pollutants over time and over large geographic areas. In areas where ground-based monitoring stations are absent, satellite-derived aerosol optical depth (AOD) and tropospheric nitrogen dioxide (NO_2) column data serve as proxies for estimating pollutant concentrations. It has shown how useful satellite data can be for epidemiological research, particularly focusing on exposure to pollutants in both rich and poor countries. However, the authors highlight issues of calibration uncertainty, retrieval and atmospheric correction errors, and other factors, which can weaken accuracy. Their study rightly concluded that satellite observations are critical for improving air quality models. Furthermore, they should be prioritized for global public health advocacy.

The evolution of inexpensive air quality sensors in Atmosphere was reviewed by Tariq, Touati, Crescini, and Manouer (2024), focusing on calibration techniques, trade-offs on accuracy, and the scalable monitoring deployment models [3]. The review mainly highlights the significance of inexpensive air quality sensors in expanding monitoring capabilities, despite some limitations in comparison to reference-grade stations. These sensors can be categorized according to the type of pollutant, the technique of calibration, and the deployment strategy of the network. Although affordability makes low-cost sensors appealing for widespread deployment, competing influences such as temperature, humidity, and long-term drift will generate biases. Correction algorithms and calibration frameworks can be used to improve the reliability of the sensors prior to their integration into the national air monitoring networks. The authors of this paper pointed out that, with the calibration and assimilation methods, air quality monitoring will be achievable in low-income countries using inexpensive, or low-cost sensors.

Malings et al. (2024) illustrated a data-fusion approach that integrates models, satellites, and ground monitors to better estimate and forecast air quality [4]. Data fusion studies indicate that combining models, satellites, and ground-based monitors improves air quality estimates and forecasts. The authors employed sophisticated data assimilation methods to combine satellite data with ground sensor data and then fed this, along with atmospheric simulations, to the assimilation model. This varied the estimate of the pollutant concentration that the model tracked across space and time. The study noted that machine learning algorithms perform better on fused datasets than on siloed ones. In addition, the paper noted that hybrid systems improve short-term forecasts and strengthen policy. The authors argued that fusion-based observations are the most effective route for comprehensive frameworks for estimating air quality indices.

Zhang et al. (2025) assessed the potential of geostationary satellite observations to improve near-real-time air quality forecasting [5]. Their transformative capabil-

ities for forecasting air quality were emphasized by research detailing the use of geostationary satellites. Unlike polar-orbiting satellites that provide infrequent and temporary snapshots of earth, geostationary satellites offer continuous monitoring of a specific area and can track ozone, sulfur dioxide and aerosol pollutants in near real-time. This work demonstrated the value of geostationary satellites in improving time coverage and developing advanced warning systems for high-pollution days. The inclusion of these data in atmospheric modeling tended to improve the accuracy of the forecasting by integrating with ground stations. However, the authors mentioned the need for further development to address problems of cloud cover, retrievals, and other inaccuracies. They also believed geostationary satellites would be fundamental to the future air quality monitoring systems.

Semlali et al. (2021) proposed SAT-CEP-monitor, a CEP-enabled architecture integrating satellite streams with event processing for AQ monitoring [6]. The SAT-CEP-monitor framework offered a new combination of a hybrid monitoring paradigm that merges complex event processing (CEP) with satellite remote sensing for the purpose of real-time air quality monitoring. The system uses multi-source data streams, which basically include IoT-enabled sensors and the satellite systems for automating pollution events, processing, and monitoring in real time. Studies have shown that this architecture improves scalability and helps to respond quickly to pollution sources. The stated goals are supported by the city ecosystem and real-time functioning. The integration of CEP and remote sensing provides the system with initial formation of pollution hotspots and automated environmental management.

Vajs, Drajic, and Cica (2023) studied ML-driven calibration propagation to improve low-cost sensor reliability within hybrid networks [7]. The literature describes more robust data-driven calibration approaches for hybrid air quality monitoring networks as an emergent strategy for improving the reliability of low-cost sensors. The authors of this work described the use of machine learning to automate real-time adjustments of reference-grade monitors to be used with distributed sensor networks, thereby considerably reducing the need for frequent re-calibrations, all the while maintaining the spatial reliability of the distributed sensor networks. The described machine learning approach shows innovation and flexibility in adapting to varying conditions of the environment. The calibration campaign offers a cost-effective scalable foundation of dependable monitoring that the authors claim will support broad use in urban and regional settings.

Ajnoti, Gehlot, and Tripathi (2024) developed optimization methods for hybrid instrument networks to maximize coverage and representativeness at minimum cost [8]. Reports have been made on the examination of the optimization strategies of hybrid instrument networks on the air quality monitoring systems focusing on the combination of ground stations, satellites, and atmospheric frameworks. This effort aims for economically sustainable supervision, optimizing the sensor's supervision by improving coverage and representativeness and decreasing overlapping and cost. This study showed that hybrid optimized networks exceeded the performance of randomly assigned networks on forecasting and precision. The comprehensive study on the topographic, meteorological, and emission source framework and other com-

binations bolstered robust monitoring. They illustrated that hybrid optimization method is effective for network development and maintaining monitoring sustainably. This is feeding predictive models that shape policy for policy direction.

There have been advances in satellite remote sensing, low-cost sensors, and hybrid fusion architectures, but there are still significant gaps. Satellite products have wide spatial coverage but have vertical-column versus surface mismatches, retrieval uncertainties, and temporal gaps which negatively affect local exposure estimates [2]. In addition, low-cost sensors require calibration, drift correction, and quality control pipelines, which are costly and time-consuming, all to obtain reference-grade accuracy [3][7]. Data fusion techniques improve spatial completeness but there are methodological challenges. For example, satellites, models, and ground monitors capture data across different spatial and temporal resolutions, uncertainty is not propagated to fused outputs, and fusion techniques do not generalize to different regions with alternative emission and meteorological conditions [4][5]. Event stream processing or complex event processing (e.g., SAT-CEP) combined with satellite streams has been proposed but there are latency, robustness, and scalability challenges for city-scale deployments [6]. Moreover, optimization of spatial networks has been designed but issues remain with optimal fixed sensor location within budget limits and the need for dynamic reconfiguration at runtime due to shifting urban activity, sensor health, and possibly control [8]. These gaps underscore the demand for standardized methods concerning calibration and transfer, uncertainty-aware fusion algorithms, and long-term field studies focused on validating hybrid systems across varying landscapes.

2.3 Machine Learning and Deep Learning in AQI Prediction

Liu, Cui, and Liu (2024) assessed the performance of classical machine learning (ML) classifiers compared to ensembles regarding AQI class prediction using a combination of pollutant and meteorological features [9]. Numerous studies have implemented diverse machine learning (ML) techniques to forecast air quality classifications based on AQI tiers, in conjunction with meteorological and pollutant data, and subsequently assessed the predictive performance of those techniques. The machine learning techniques examined included support vector machine, random forest, and gradient boosting, which were compared to more advanced predictive deep learning techniques. In terms of the best results of accuracy, precision, and recall came from ensemble methods and especially random forests. Their research further strengthened the predictive value of the algorithms. The models effectively used time-series data of pollutants ($PM_{2.5}$, PM_{10}, O_3, and NO_2) and captured temporal trends that closely matched ground data. Research has shown that machine learning approaches could be effectively utilized for discrete classification of AQI categories, and ensemble learning has been suggested as a strong baseline model for AQI classification systems in the future.

Saleh, Abohamama, and Tolba (2025) designed an intelligent real-time AQI classification system for smart-home digital twins using IoT sensors and ML [10]. These systems use IoT-embedded AQI sensors and machine learning classifiers to provide real-time air quality class awareness both indoors and outdoors. This study emphasizes the importance of locally operated lightweight systems that provide real-time notifications to residents. The system achieved AQI classification with a high degree of computational efficiency, making it a viable candidate for embedded system implementation. This study demonstrates that IoT real-time AQI classification improves user awareness and proactive environmental control at the household level.

Mehmood, Rehman, and Choi (2025) introduced Vision-AQ, a multi-modal DL framework combining imagery and sensor data for localized AQI classification [11]. Vision-AQ is a multi-modal deep learning framework that classifies localized air quality indices in smart city environments. It merges image-based pollution indicators and traditional sensor data, utilizes convolutional neural networks, and implements attention mechanism techniques to model spatial and temporal dependencies. This outlines how interpretability of Vision-AQ by using various visualization tools showed how various techniques used impacted the decision made during classification. Moreover, the model showed considerable improvements in functionality and efficiency than classical models. The authors fully implemented the integration of visual and numerical data for the tracking and classification of urban AQI in Vision-AQ.

AirNet is an Air Quality ML forecasting system, developed by Rahman et al. (2024) with a Django web interface that seeks to offer real-time alerts across a wide geographical area [12]. The work integrates a substantially large dataset (public data covering thousands of cities) and multiple classifiers, such as Random Forest (RF), Decision Tree, SVM, Logistic Regression, KNN, Linear SVC, and Naïve Bayes, following thorough data preprocessing (missing value treatment, label encoding, feature correlation, and analysis). The Authors state that tree-based methods (Random Forest and Decision Tree) yielded the best predictive results, thus powering the web dashboard for real-time AQ classification and alerting. AirNet is focused on user-friendliness by exposing the outputs from the classifiers and targeting two main objectives: (i) high short-term classification accuracy for AQ categories and (ii) operational readiness for public health alerts. The Authors provide the inconvenience of dependency on public data quality, regional Ara biases, and on-site sensor calibration and suggest further research on the generalizability of the models to under-sampled regions.

In 'SMOTEDNN: A Novel Model for Air Pollution Forecasting and AQI Classification', Haq (2021) addressed the problem of class imbalance by implementing a synthetic oversampling method integrated with deep neural network frameworks [13]. Through the utilization of the synthetic data grown by SMOTE, the authors were able to 'grow' the smaller AQI categories for classification purposes on a deep neural network. The evaluation metrics indicated a profound improvement on the performance of the recall and F1 scores on the minority classes. Most of the advances were in the higher-risk categories of pollution. This study revealed significant improvements in the overall classification accuracy and lack of deep learning and

oversampling techniques on AQI datasets.

Kutala et al. (2024) in 'Hybrid Deep Learning-Based Air Pollution Prediction and Index Classification Using an Optimization Algorithm' presented a CNN-enhanced model optimized through evolutionary search for faster, more precise AQI categorization [14]. This approach aims to achieve a balance of model performance with feature extraction as suggested by the use of evolutionary optimization in hyperparameter tuning. It was noticed that the hybrid framework based on CNN significantly surpasses the predictive power of conventional designs in classifying AQI and was designed with more advanced means of optimization to achieve a reduction in training duration and resource utilization.

Jalali et al. (2025) employed a hybrid method in 'Scalable AI-driven Air Quality Forecasting and Classification for Public Health Applications'. Ensemble methods like, Random Forest (RF) and XGBoost and deep learning techniques, as an example, LSTM, CNN, and TSMixer were combined for socioeconomic and geographically contextualized classification and regression modeling of Air Quality (AQ) values [15]. This work sets a new standard for integrated machine learning and public health applications within economically resource-limited settings. Using a novel data augmentation approach, the study clusters and associates cities with comparable pollution characteristics. The author combines SHAP with LIME to explain model predictions, and provides a Django app as real-time access to the model. Random Forest, for classification, achieved an astonishing 99.96% accuracy (slightly better than XGBoost), while TSMixer was remarkable in regression with R^2 at 0.9861 and MSE at 0.0278. The emphasis is placed by the authors on the fact that explainability and operational readiness were central to their design: it is confirmed by SHAP analyses that NO_2 and PM_{10} consistently ranked among key drivers, and fast inference (≈ 0.0289 s for ensembles) was prioritized by the deployment architecture. While deep models like CNN/LSTM demonstrate strong predictive capabilities, the paper notes that their high computational cost – up to about 1.25 seconds per prediction – can limit real-time performance in constrained environments.

'Air Quality Index Classification for Imbalanced Data Using Machine Learning Approach' by Jayadi, Lauro, Rusdi, and Handhayani (2024) demonstrated effective rebalancing of skewed AQI datasets in Indonesia using ensemble and resampling methods [16]. Hidayat and coworkers investigated AQI classification in Indonesia, whose study primarily concerned the use of machine learning on imbalanced datasets. They assessed the comparative performance of classifying AQI with SVM, Random Forest, and deep learning models while employing data resampling to balance the training datasets. Here, the authors observed a thing. Despite the extreme class imbalance, machine learning, particularly Random Forest as a baseline, was capable of reliable AQI classification. The importance of data preprocessing in developing countries was pointed out by the authors, as AQI datasets are often unbalanced and incomplete.

Rezk et al. (2024) proposed SFS-driven feature selection combined with ML algorithms to produce sustainable, efficient AQ detection systems [17]. A robust model for air quality detection using sequential forward selection (SFS) for feature selec-

tion and various machine learning methods for classification. The model achieved significant classification performance while maintaining low data dimensionality by modestly detecting features such as $PM_{2.5}$, temperature, and wind speed. The results revealed that the implementation of SFS before classification contributed positively to the interpretability and efficiency of the models, with Random Forest and Gradient Boosting exhibiting the most prominent results. The study pointed out the significance of feature engineering in the development of sustainable and resource-efficient systems for predicting the AQI.

Zhang, Zhai, Kong, and Zhang (2025) combined stacking ensembles with SHAP analyses to both improve AQI classification and explain feature impacts [18]. To advance the AQI classification, integrated stacking ensemble learning, and SHAP-based explainability. In an effort to optimize accuracy within a stacking architecture, the framework integrated several classifiers, including gradient boosting and deep neural networks. The authors were able to justify model predictions at both global and local levels through SHAP analysis, illustrating the impact of various weather and pollutant variables on AQI levels. The findings indicated that interpretability is a feature that can be integrated within high-performance ensemble classifiers without significantly losing accuracy. This can create a valuable linkage between prediction and transparency.

Kalantari et al. (2025) compared shallow and deep learning methods for AQI forecasting, clarifying trade-offs between complexity and data requirements [19]. They investigated the relative strengths of shallow learning compared to deep learning for AQI forecasting and classification, and performed a comprehensive range of model evaluations starting from logistic regression, decision trees to CNN, and recurrent neural networks (RNNs). He noted that although shallow learning models are faster to train and less resource-consuming, deep learning models unequivocally surpass shallow models, particularly in the learning of spatio-temporal dependencies as well as the nonlinear dynamics of pollutants. Importantly, the authors emphasize that context is important when considering which models to use. Shallow models are fast and resource-efficient, whereas researchers require deep learning for high-accuracy urban deployments, which is the primary goal of large-scale AQI research. It understands the trade-off between computational resources and achieving predictive performance in AQI research, and that's why it is a valuable contribution.

Maciejewska, Azizah, and Szczurek (2024) in 'IAQ Prediction in Apartments Using Machine Learning Techniques and Sensor Data' constructed a framework designed to predict indoor air quality using readings from multiple sensors located in data-equipped residential settings [20]. Using algorithms like Random Forest, Gradient Boosting, and Support Vector Machine, the authors predicted the concentrations of several pollutants: CO_2, VOC, and $PM_{2.5}$. The findings showed that ensemble-based models surpassed baseline classifiers in accuracy and offered interpretable feature importance for real-time applications with the Random Forest model. The authors noted the potential for incorporating machine learning-based indoor air quality prediction features into an automated smart home system and an automated public health alert system.

Verma and Kumar (2022) in 'Machine Learning-Based Time-Series Models for Effective CO_2 Emission Prediction in India' extended time-series learning methods applicable to broader AQ forecasting and environmental management domains [21]. The seasonal and trend-based emission variation was captured using shallow learning approaches like ARIMA and SVR alongside ensemble techniques in an attempt to forecast. The research found that machine learning techniques were much more accurate in forecasting emissions, unlike the traditional statistical methods, especially when applied with multiple variables like energy consumption, traffic flow, and meteorological variables. Integration of the predictive models into wider systems of air quality management along with predictive model integration into systems of air quality management will aid in the advancement of environmental management as well as the air quality regulatory frameworks within the study area. Highlighting the model selection process, worksheet engineering, and hyperparameter optimization reveals the authors' dedication to crafting practical and sufficient frameworks. These can be embraced by researchers and practitioners in the fields of sustainable computational and environmental monitoring.

In the broader methodological literature on hybrid deep learning and optimization as well as CO_2 time-series work, constructive modeling building blocks are provided, but questions of generalizability, reproducibility, and alignment with the target domain remain. While hybrid deep learning and optimization pipelines deliver strong predictive performance, they are often clouded by excessive sensitivity to hyperparameter tuning, the lack of open, versioned code and datasets, and thus, reproducibility and comparative studies across the literature [20]. The time-series models designed for CO_2 or emissions control provide transferable techniques. However, pollutant dynamics, such as multi-pollutant coupling, episodic events, and other dynamics create statistical challenges requiring more specialized and often inconsistently used architectures, assessment frameworks, and metrics [21]. Lastly, there is a systemic need for establishing community standards (e.g., benchmarks, preprocessing recipes, unified metrics for misbalanced AQ classification) and there is a systemic need for additional public datasets containing metadata (e.g., type of sensors, calibration history) so that models trained by one group can be adequately assessed and implemented by others. From a methodological standpoint, addressing these issues will foster the advancement of AQ ML systems that are dependable, comparable, and ready for real-world application.

Class imbalance concerning the 'good' AQ categories leads to lower detection rates of the minority class (hazardous levels). Oversampling techniques and weighted losses can help but can also cause the methods to be overfit or create unstable decision boundaries [13][16]. While many studies examine multiple learners (SVM, RF, XGBoost, LightGBM, CNN/LSTM ensembles) using local datasets, there is a lack of standardized and multi-city benchmarks, and the uneven approach to preprocessing (imputation, feature engineering, lag selection) leads to unreliable cross-paper comparisons [9][17][19]. These deep multimodal approaches (e.g., Vision-AQ) enhance spatial discrimination but also increase model complexity and inference costs, hindering on-edge or real-time deployment [10][11][12][14][15]. Most classification studies lack explainability and uncertainty, making it hard for stakeholders to trust or act on the AQ automated category alerts [18]. The issues of transferability (when

models trained in one city do not work in another) and a lack of longitudinal assessments mean that many classification studies do not sufficiently demonstrate the non-impact of seasonality, concept drift, and sensor aging. These issues underscore the need for unified benchmarks, rigorous imbalance-aware validation, lightweight deployable explainable classifiers, and systematic studies on transferability.

2.4 Explainable AI (XAI) Techniques – SHAP for AQI and Pollutants

A hybrid random forest–ARIMA forecasting model with SHAP-based interpretation for AQI drivers was presented by Yenkikar, Mishra, Bali, and Ara (2025) [22]. The researchers aimed to improve ARIMA's weaknesses by combining it with tree-based machine learning techniques. SHAP was used to determine the contributions of individual pollutants and meteorological parameters like $PM_{2.5}$, PM_{10}, and others on the classification of the AQI. Compared to ARIMA and Random Forest on their own, the ARIMA–Random Forest hybrid model was particularly accurate in short-term forecasts of AQI. SHAP uncovered that $PM_{2.5}$ was the dominant predictor and most influential of all pollutants in every time period, and that temperature and wind speed influenced AQI differently in different seasons. The model's qualitative predictions of smogs in winter and washes of pollutants in the monsoon season demonstrated SHAP's explanation of the model and the environmental phenomenon. The balance of predictive power and transparency of this work is the most valuable aspect, providing forecasting to environmental policymakers with confidence that they understand the causal factors.

XGBoost with SHAP was leveraged by Andenna, and Gagliardi (2024) in 'Insights into Air Quality Index Variability with Explainable Machine Learning Techniques' to map pollutant dependencies and to assess feature relevance across various cities [23]. Using explainable machine learning to study AQI variability in different urban areas, the framework used XGBoost for classification and used SHAP to measure the impact of the features. Considering urban form predictors, the analysis of socio-economic variables provided a nuanced understanding of the factors influencing the AQI. The analysis revealed that $PM_{2.5}$ and NO_2 continued to dominate the important features in most areas. However, SHAP exhibited certain contextual variations; for example, in megacities, traffic density was a major contributor, and in peri-urban areas, agricultural emissions were the most important. Contextual analysis through SHAP within a multi-city comparative framework offers a better understanding of the variation in AQI. The significance of this contribution lies in establishing policy context explainability and informing actionable policy from targeted vs. one policy fits all. The limitations address the data availability for the cities and the inconsistencies that may come with it. Nevertheless, the value of this research lies in establishing a precedent for multi-dimensional AQI analysis using SHAP.

Iffadah, Trimono, and Prasetya (2024) applied SHAP to interpret XGBoost predictions of Jakarta's air quality, highlighting dominant pollutant drivers and seasonal

effects [24]. Here it is used to explain the XGBoost-based air quality prediction models in Jakarta and gain an understanding of the interaction of pollutants in densely populated areas. While the classifier did its job of categorizing AQI levels, the SHAP approach was crucial for understanding the model, as it explained the level of attribution in the various features. It was able to identify $PM_{2.5}$ and CO as the key contributors of days with a poor AQI and showed how meteorological elements like rainfall helped to restore deteriorated air quality. The seasonal effects, as SHAP explained were demonstrable, specifically the atmospheric pollutants increases during the dry months when the atmosphere is less cleansed and stagnant. This kind of local interpretability is indispensable for decision-makers who are looking for tailored, region-specific insights. The authors further explained that the lack of explainability offers no trust in AI systems for AQI management, particularly in situations where public access to management is life changing. This study enhances the literature by localizing SHAP interpretations.

'Machine Learning-Driven Framework for Realtime Air Quality Assessment and Predictive Environmental Health Risk Mapping' was presented by Rajesh, Babu, Moorthy, and Easwaramoorthy (2025), incorporating real-time AQ prediction along with SHAP explainability that is linked to public health indicators [25]. Their model combined ensemble methods such as XGBoost and Random Forest with health risk indices, producing not only AQI classification but also mapping of exposure risks for vulnerable populations. SHAP analysis has specified the most salient pollutants concerning health effects, with $PM_{2.5}$ and ozone repeatedly cited as the most important drivers of respiratory illness risk. A novel contribution of this study is linking the importance of pollutants as determined through SHAP with epidemiological outcomes, thus integrating environmental informatics and public health. Real-time mapping further showcased the framework's utility for rapid urban response. Although operational compute power was identified as a constraint for wide-ranging applications, the approach demonstrates the use of interpretability tools, in this instance SHAP, for purposes beyond mere classification to health-oriented actionable frameworks.

Das, and Gupta (2025) developed an AQI prediction model that uses SHAP to explain local and global feature contributions in classification outputs [26]. For AQI classification, the framework used ensemble learners, while SHAP was applied to ascertain the contribution of features at the global and local levels. Unlike previous work focusing disproportionately on pollutants, this study broadened the feature set to include meteorological variables, such as humidity, solar radiation, and wind direction, for a richer interpretability framework. SHAP results showcased why $PM_{2.5}$ occupied the highest rank in global feature importance. On the other hand, localized forecasts were more dependent on the input variables, which typically were the localized weather variables. This signifies the interdependence of weather and localized pollution. Basically, this illustrates the importance of regional policies, in the context of the demonstrated SHAP explanation of context drivers of AQI. Furthermore, case studies were utilized by the model in which the recorded instances of smog converged with the SHAP explanations, affirming the applicability of the model in real life. While the high-quality input data limitation is real, the value of the work is in the interpretability SHAP provided, which is a first step toward

operationalizing interpretability in AQI monitoring systems.

Chandrashekar et al. (2025) integrated physical modeling and deep learning in 'A Hybrid Physics-Informed Neural and Explainable AI Approach for Scalable and Interpretable AQI Predictions', enhancing scientific interpretability alongside predictive accuracy [27]. This innovation introduced the theory of the atmosphere within the deep learning paradigm and, therefore, decreased the data-driven reliance and overfitting risk. The SHAP method was used in predicting feature importance, thus elaborating on the interface of domain knowledge and machine learning. The hybrid model provided the authors with more accurate results for predicting extreme AQI events and overall outperformed traditional Deep Neural Networks. Significantly, SHAP explanations were consistent with physical models, illustrating how inversion layers and wind stagnation were vital to worsening pollution. The consistency between physics-informed modeling and SHAP-based interpretability fosters trust towards AI predictors from domain experts. The integration of physics and AI model SHAP's validation to hybrid models as the central focus of the paper's contribution.

Wu, Chen, Li, Wang, and colleagues (2024) in 'A Novel Framework for High Resolution Air Quality Index Prediction with Interpretable Artificial Intelligence and Uncertainties Estimation' developed an explainable AQI deep learning predictive model [28]. High-resolution spatial, meteorological, and emission factor data are combined by this model, and SHAP values are applied to assess variable impacts across places. The ability to separate predictive confidence from data-related noise helps to increase the confidence of the estimates. Therefore, the reliability for operational contexts is strengthened. This study shows how mixing uncertainty estimation with interpretability contributes to increasing policy confidence and applicability towards AI-driven predictions of air quality.

SHAP and other XAI techniques have become mainstream for de-blackboxing AQ models, but there are still numerous methodological and practical issues. First, in operational systems, real-time explanation is often infeasible because SHAP explanations for large ensembles and deep networks become computationally expensive [22][23]. Second, post-hoc local attributions are provided by SHAP but may be unstable under correlated features or distributional shift; this complicates the interpretation of variable importance for policy decisions by practitioners [24][27]. Finally, the integration of uncertainty quantification in explanations (i.e., to what extent a user can trust a SHAP attribution if the model is likely wrong) is still largely unexplored, even though several studies apply SHAP to individual instances or features. Although a couple of studies aim at merging SHAP with uncertainty quantification, the field still lacks standardized approaches [25][27][28]. Fourth, the limited evaluations that look into the impact of SHAP value outputs on decisions made by domain experts or on alert explanations made to the public are addressed in the literature. This literature points towards the need for more rapid and stable XAI integrations to be utilized for streaming AQ systems, protocols that explain and incorporate uncertainty as well as user studies to ascertain the relevance and integration of the explanations into public health and policy workflows.

2.5 Smart Campus / Student-Centered Solutions

Ramirez et al. (2023) reviewed campus AQ monitoring systems and impacts, documenting sensor mix, network design, and campus health impacts [29]. These systems acquire and considerably help in the determination of the health policies, and help with the education of the student population and the staff. They assist in the determination of the spatial and temporal distributions of pollutants, $PM_{2.5}$, NO_2, and O_3. Provided sensor deployment frameworks to capture the spatial and temporal dynamics of pollution, sensor drift, maintenance, and weather for the calibration, processing, and validation of active readings. The review documented the quality of interdisciplinary research that students gain from being active participants in the air quality studies on campus, enabling learning and teaching in environmental monitoring, data science, public health, and interdisciplinary research. The value of synergizing the adaptive management of air quality monitoring systems and the campus-wide sustainability interventions was also demonstrated, such as changing the levels of pollutants to control the movement of vehicles, greening areas, and running specific awareness campaigns. This study then underscored the monitoring of air quality in a university environment for research and teaching purposes.

Kozłowski et al. (2025) detailed 'Autonomous System for Air Quality Monitoring on the Campus of the University of Ruse: Implementation and Statistical Analysis', that effectively deploys multi-sensor networks for real-time alerting and continuous data collection [30]. The system provides real-time analysis and reports on the pollutants $PM_{2.5}$, PM_{10}, CO, NO_2, and other meteorological parameters, such as temperature and humidity. Sensors organized the collected pollutant data in a real-time cloud-based system with visualization, automated alerts, and time series analysis for various predetermined parameters. The focus of the authors addresses the autonomy of the system. Machine learning algorithms helped to fix the drift of the sensors and to increase the accuracy of predictions concerning the concentration of pollutants. Microclimate investigations showed the extent of localized pollution hotspots and were the basis for mitigation action. The system's combination of intensive monitoring aided by the automated analytics permitted the formulation of policy recommendations based on the system's real-time environmental management dashboards. The potential for autonomous monitoring frameworks to be rapidly adopted across smart campus projects was described in the research, and the implications of this for student well-being, administrative actions, and the planners of eco-friendly campus facilities are considerable.

Yadav et al. (2020) demonstrated deep transfer learning with satellite imagery for urban AQ estimation, showing transferability across cities [31]. Estimation of air quality on campuses utilizing methods described in the literature for urban settings and adapting satellite imagery stems from deep transfer learning conducted. The application of Convolutional Neural Networks and Transfer Learning with satellite images to calculate Aerosol Optical Depth and neural network pollutant concentration noted the integration of satellite images, which provided continuous coverage, with ground-based, handy, and sparse monitoring systems. Data fusion techniques,

which predicted the AQI for $PM_{2.5}$ and NO_2 and used meteorological and topographical variables for precise estimations described. While the air quality study had a primary focus on city-scale air quality assessments, the method used is still viable for air quality assessments on university campuses, where satellite data could augment the sensor networks for comprehensive assessments. The study demonstrates the importance of predictive analytics, which allows pre-warnings and action plans to be developed to reduce students' exposure to high pollutant concentrations, showing the potential of such analytics in a campus context. The researchers mainly highlighted the importance of combining deep learning (DL) and remote-sensing techniques to improve the consistency and fine-scale measurement of air quality in the low-resource university campuses.

Yang, Hu, Bian, and Song (2019) introduced ImgSensingNet, a UAV vision-guided aerial–ground sensing system to enhance spatial AQ observations [32]. The system used UAVs with multispectral cameras and low-cost sensors for the dynamic measurement of PM2.5, PM10, and other gaseous pollutants, as well as for air quality assessments in campus-like environments. Data from UAVs and terrestrial sensors were integrated to create a multi-dimensional spatial profile of air pollution, and the UAVs were also used to fill data collection gaps of terrestrial sensors. The system design, integration of datasets, and airborne data acquisition to enable the closure of gaps in continuous monitoring and acquisition streams were detailed in the paper. The authors showcased the vital role of aerial data in addressing gaps in monitoring coverage caused by sensor networks. The integration of UAVs with predictive analytics unveiled dynamic shifts in the pollutants and the impacted microclimate. The research provided some practical applications to environmental management, such as emission hotspot identification and appropriate mitigation for student-centered spaces. The opportunity to integrate UAV sensing with school curricula, where students would collect, analyze, and interpret air quality data, was also noted by the authors.

ML techniques were applied to PM2.5 forecasting on a university campus by Panaite et al. (2024), with improved local forecasting for student health being demonstrated [33]. Machine learning techniques were utilized for forecasting the levels of PM2.5 based on historical data for air quality prediction and improvement on the Petroșani University campus. Random forest trees were created, then gradient boosting methods were developed, and also the support vector regression (SVR) models were built based on prior data that included weather, emissions, campus activities, and pollutants. Predictive data dashboards were created for real-time decision-making and recommendations about outdoor traffic and activities. The importance of tuning the models and selecting the right features is illustrated with local emissions trends and campus microclimates in mind. Results from this research advocate the implementation of data-driven strategies in support of campus activity policies, enhancing sustainability while providing a higher margin of health safety for students and staff. Predicting accuracy was higher in the research models as opposed to predictive accuracy compared to older methods which relied on pure statistics.

Godasiaei, Ejohwomu, Zhong, and Booker (2025) authored 'Integrating Experimental Analysis and Machine Learning for Enhancing Energy Efficiency and Indoor Air

Quality in Educational Buildings', effectively linking building design and predictive analytics to optimize health outcomes [34]. This study integrated concept-driven adaptive control with the optimization of HVAC systems, coupled automated control of the systems and robust measures to control the indoor pollutant concentrations. Core machine learning models on occupancy and environmental data with respect to $PM_{2.5}$, CO_2, and VOC predicted energy-efficient operation strategies within the control measures scheduled during the periods of pollutants concentration control. The authors documented the benefits of adaptive intelligent energy monitoring that balanced dynamic changes in building use and occupancy, and argued the energy optimization control system could function in a managed energy adaptive mode. The study also documented the added value of environmental education as the students developed systems for analysis and evaluation. The research described at the conference demonstrated that the real-time monitoring and control of dynamic energy systems on the campus combined with machine learning provides a powerful, albeit untapped, instrument for managing the quality of indoor environments.

A modular IoT sensing platform for air quality monitoring and data fusion using hybrid learning was developed by Sridhar et al. (2022), incorporating low-cost sensors and machine learning (ML) to detect $PM_{2.5}$, CO_2, VOCs, and other pollutants [35]. The design of the modular IoT platform provides the ability to scale and customize campus deployments to independently monitor specific buildings and outdoor areas. Predictive performance has been improved through a hybrid learning approach by integrating historical data and real-time information. This study basically focused on the importance of actionable insights, thus providing stakeholders that included campus admins and students, with dashboards and automated alerts. Integrating sensor technology with analytical intelligence has expanded environmental management capabilities and has helped the system support health and sustainability efforts. The system's monitoring and learning capabilities serve dual purposes. The authors highlighted the educational value of the system when having students perform calibration of sensors, interpret data, and participate in collaborative activities of consequential decision-making.

Health awareness, as seen in student-centered AQ deployments, also points to positive practical implications. However, these studies are also limited in the evaluation of the health impacts and the scaling of the studies. Most campus studies tend to focus on sensor deployments and local machine learning forecasting for $PM_{2.5}$ or indoor IAQ in isolation of more extensive student health behavior or outcome studies that would shift or demonstrate changes in student health behavior or public health alert systems [29][34]. System design also tends to neglect user privacy and data governance, especially pertinent when personal or indoor sensors are deployed to a population of students [29][35]. Integration of indoor and outdoor AQ models, model adaptations for building-level ventilation, the design of low-complexity classifiers to be run on edge devices while meeting the accuracy needs of the classifier and other such elements are also substantial technical gaps that need to be addressed [30][35][34]. Additionally, existing campus pilots tend to be single-site and short-term. There is a lack of multi-campus comparative studies and controlled evaluations, such as randomized information interventions. To address these requirements, we need longitudinal, multi-site deployments covering measurements

of the technical and human dimensions along with purpose-built privacy controls, lightweight explainable algorithms designed for edge and dashboard delivery, and explainable models tailored to edge and dashboard delivery.

Chapter 3

Methodology

3.1 System Overview

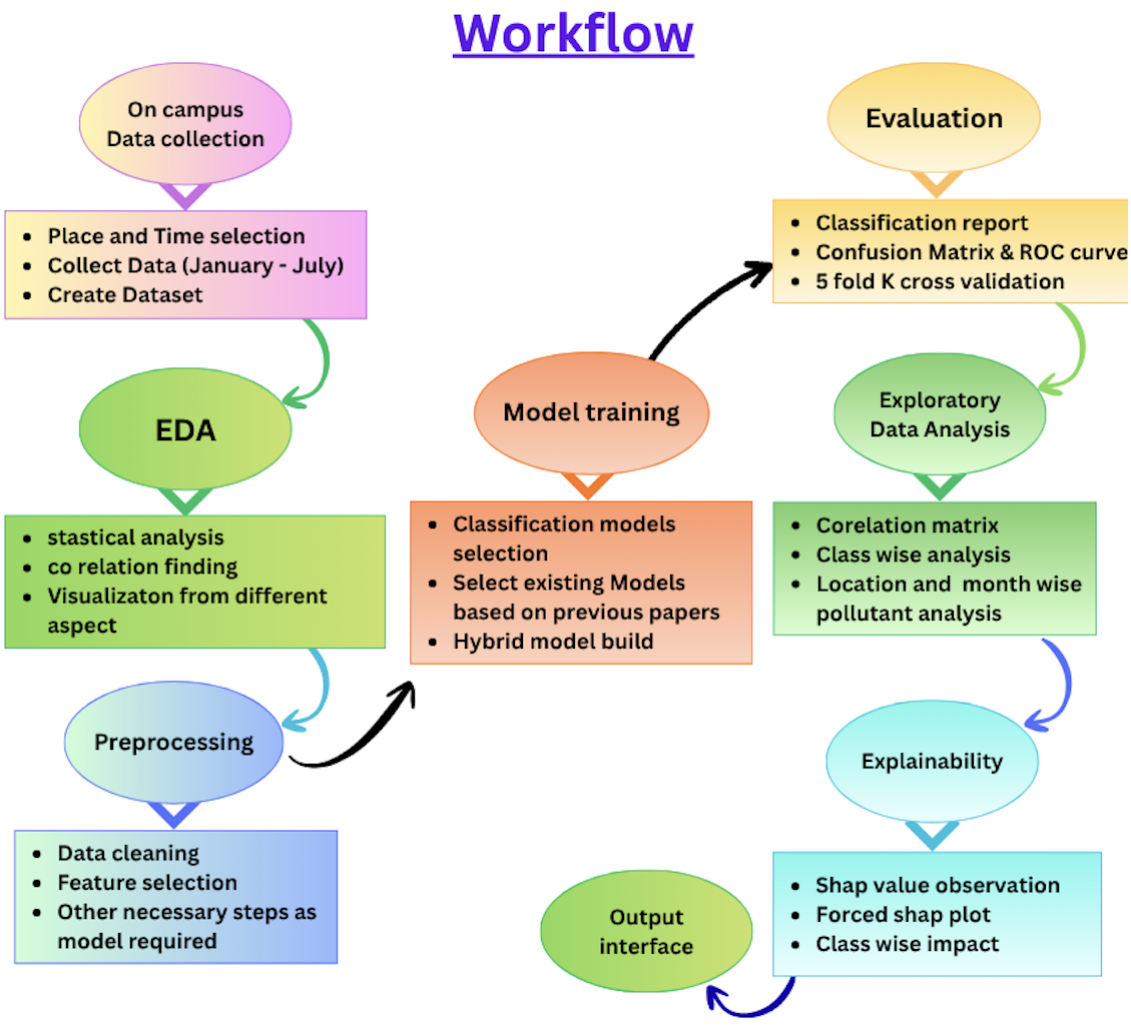


Figure 3.1: System Overview.

The root of this study is collecting data from university campus. As a consequence, started collecting data from selective places on the campus at specific times and enlarging the dataset.

This study also has Exploratory Data Analysis (EDA) for clear statistical understanding of AQI with the pollutants as well as location-wise, month-wise, and other analysis with proper visualization.
After preparing the dataset, it needs to be processed for the main branch of research work. As part of preprocessing, the dataset needed to clean, selected pollutants as features, and took other necessary steps for further work.
Moreover, after finishing those necessary steps, trained classification machine learning and deep learning models on the dataset and also test for the target variable, AQI classification level prediction. Additionally, in this study, a new hybrid model was build with a combination of machine learning and deep learning.
However, as an evaluation of trained and tested models, classification report was used for each class's accuracy, precision, and recall values as well as 5-fold k-cross validation. In addition, this study uses a multi-class confusion matrix and ROC-AUC curve as a part of the evaluation.
Furthermore, for better understanding of pollutant feature patterns and contributions for hybrid model predictions, we observe SHapley Additive explanations force plots as a part of explainability.

3.2 Dataset and Features

Device Description:
As an innovative monitoring device with integrated measurement modules, the MT15-Tuya multifunctional air quality monitor has the specialized ability to sense volatile organic compounds. Without the assistance of an external phone, application, or even the Internet, the device in real time detects gas and particulate matter from the ambient air and detects pollutants as the device turns on.
The MT15 device functions as a multitasker, as it can monitor and visualize the total amount of Air Quality Index (AQI), volatile organic compounds (VOCs), formaldehyde (HCHO), particulate matter (PM), namely $PM_{0.3}$, PM_1, $PM_{2.5}$, and PM_{10}, the total carbon, and the CO_2 in the air in addition to the total carbon monoxide (CO). Moreover, it also measures environmental conditions like humidity and temperature. The device has additional timers and alarm functions to aid its usability and its LCD to visualize and monitor real-time time, alerts, and system readings. Because of the integrated approach, the device's effective multifunctional real-time monitoring systems make the MT15 perfect for an indoor as well as semi-indoor research environment.

3.2.1 Data Collection Procedure and Protocols

To study this, MT15-Tuya devices were deployed at various locations across the university to capture spatial variations in air quality, and the monitoring points were selected to represent a mix of functional areas, including classrooms, high-traffic areas, enclosed spaces, and outdoor-sensitive areas. All data collected was systematically transferred to a secure local server. To ensure data integrity, it was backed up daily. To support thorough analysis, other parameters, as an example, temperature and humidity, were, along with the pollutants, captured and saved during the

set interval. In addition, to support accurate analysis, scheduled calibrations were undertaken with reference-grade devices.
To understand how air quality changes within different zones of the university's campus, a rich variety of data was collected from numerous indoor and outdoor settings. The selected sites included the Washroom (6th floor), Inside kitchen, Fire exit (6th floor), Pond area, Library, Level-6 (inside lift and outside lift), Level-7 (inside lift and outside lift), Level-8 (inside lift and outside lift), Level-9 classroom (9E-20T), Rooftop (13th floor), Basement 1, Basement 2, and Basement 3, 4th floor, 10th floor, Auditorium. At each location, the device was positioned and measurements were taken systematically.

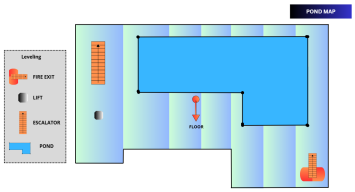


Figure 3.2: Pond.

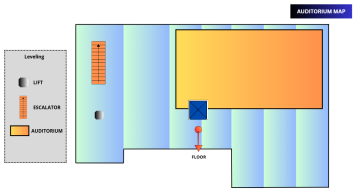


Figure 3.3: Auditorium.

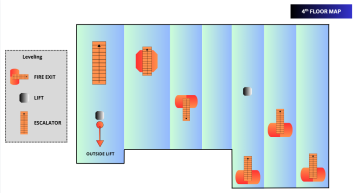


Figure 3.4: 4th Floor.

At the pond (3.2), the placement of the device occurred in front of the water body in order to facilitate the recording of outdoor environmental conditions. At the auditorium (3.3), measurements were taken outside the main entrance door to collect sample data from a high-traffic gathering area. On the 4th floor (3.4), the device was positioned outside of the lift to capture measurements of a corridor with a lot of people in the area.

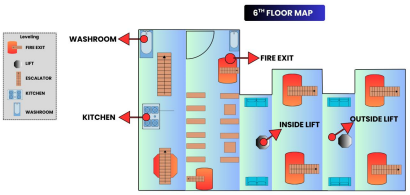


Figure 3.5: 6th Floor.

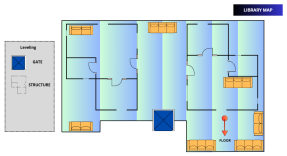


Figure 3.6: Library.

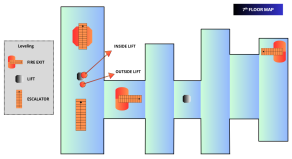


Figure 3.7: 7th Floor.

At the 6th-floor locations (3.5), the device was used in multiple spots: inside the washroom to capture air quality in a confined space, in the kitchen to assess emissions from cooking activities, inside the lift cabin to record conditions in an enclosed area, outside the lift in the corridor to compare with the surrounding environment, and at the fire exit to monitor conditions in the passageway. In the library (3.6), measurements were taken inside the floor area to represent a high-occupancy academic environment. On the 7th floor (3.7), air quality or the sample data were captured in both inside the lift cabin's and outside in the corridor's to compare air quality between enclosed and open spaces.

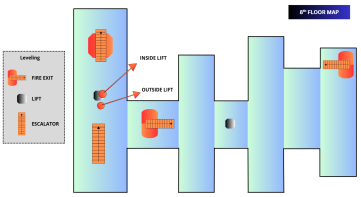


Figure 3.8: 8th Floor.

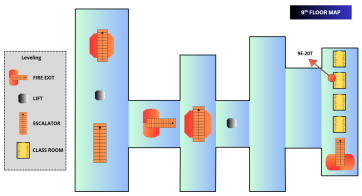


Figure 3.9: 9th Floor.

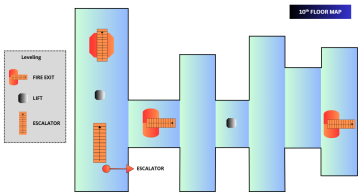


Figure 3.10: 10th Floor.

On the 8th floor (3.8), air quality was measured both inside the lift cabin and then outside in the corridor to compare conditions between the enclosed and open spaces. In the 9th-floor classroom "9E-20T" (3.9), sample data were collected inside of the classroom in a high-occupancy academic setting. On the 10th floor (3.10), just outside the lift corridor, air quality data were collected in order to assess the conditions in the common corridor area.

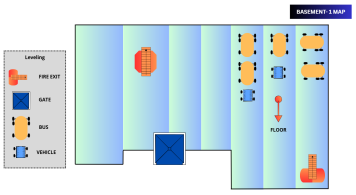


Figure 3.11: Basement-1.

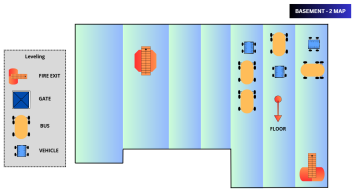


Figure 3.12: Basement-2.

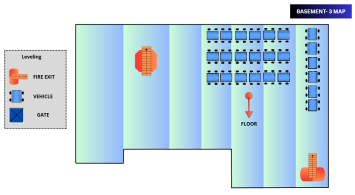


Figure 3.13: Basement-3.

In Basement-1 (3.11), air quality measurements were taken in front of vehicles and buses to capture pollutant levels in a vehicular area. In Basement-2 (3.12), the device was similarly placed in front of vehicles and buses to record air quality in another parking section. In Basement-3 (3.13), data were collected in front of vehicles to assess pollutant concentrations in the parking area.

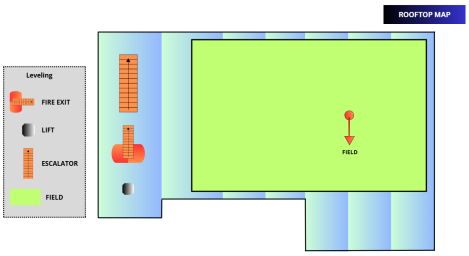


Figure 3.14: Rooftop.

On the rooftop (3.14), air quality measurements were taken in the green field area to capture outdoor environmental conditions above the building.

3.3 Data processing

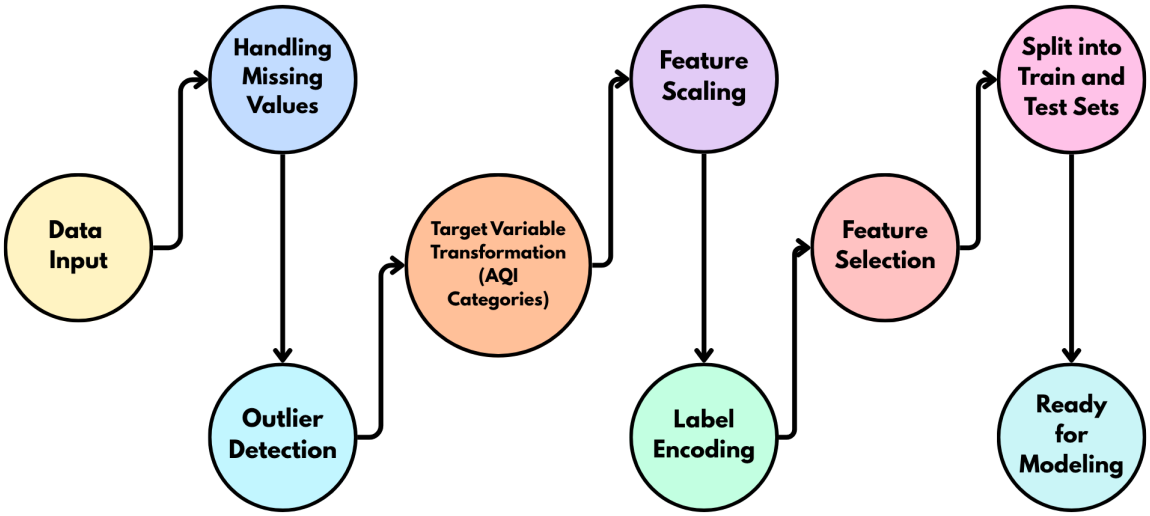


Figure 3.15: Data preprocessing

Data Input

- The dataset was imported into Python using Pandas library.
- Initially the dataset was checked for the structure, dimensions, and rarity of entries across all columns, as well as identifying missing or null values.
- Missing values in the Date and time columns were found, but sincere are not using these features that's why they are not removed.

Handling Missing Values

- Records missing data or time were deleted because they are critical for temporal analysis.
- Pollutant values which have invalid entries were converted into numeric forms.
- Any remaining missing values were covered with the average values of that feature so that we don't have to delete the whole row, less data was lost.

Outlier Detection
Outliers were identified in each pollutant feature using the Interquartile Range (IQR) method. The lower and upper bounds were computed as:

$$\text{Lower Bound} = Q_1 - 1.5 \times IQR, \quad \text{Upper Bound} = Q_3 + 1.5 \times IQR$$

The number of outliers in each feature was counted. All the outliers were kept because, as extreme pollutant values may display real word pollution spikes rather than noise.

Target Variable Transformation (AQI Categories)
The continuous Air Quality Index (AQI) values were converted into categorical labels using standard threshold levels, as shown below:

- **Good:** $AQI \leq 50$
- **Moderate:** $51 \leq AQI \leq 100$
- **Caution:** $101 \leq AQI \leq 150$
- **Unhealthy:** $151 \leq AQI \leq 200$
- **Very Unhealthy:** $201 \leq AQI \leq 300$

This transformation converted the regression task into a multi-class classification problem.

Feature Selection
Seven pollutant-based features were selected as predictors, representing both gaseous and particulate matter sources:

- CO_2 (ppm)
- TVOC (mg/m^3)
- HCHO (mg/m^3)
- $PM_{0.3}$ ($\mu g/m^3$)
- $PM_{1.0}$ ($\mu g/m^3$)
- $PM_{2.5}$ ($\mu g/m^3$)
- PM_{10} ($\mu g/m^3$)

Label Encoding
Using Lebel encoders the AQI category names were changed into integer values, so that the ML models can understand and use them easily.

Feature Scaling

- The input features were scaled using the StandardScalar that's why they have the mean value of zero and standard deviation of one.
- We know tree-based models do not need scaling, but it was done to make the features more suitable across different algorithms.

Train-Test Split

- The dataset was divided into 80% for training and 20% for testing.
- Stratified sampling was used to keep the same stability of AQI categories in both parts.

3.3.1 Exploratory Data Analysis (EDA)

Table 3.1 shows summary statistics of pollutants and AQI.

Table 3.1: Summary Statistics of Pollutants and AQI

Pollutant	Min	Max	Median	Mean	Standard deviation
CO_2 (ppm)	0.001	1139	400.0	403.041	54.513
TVOC (mg/m^3)	0.001	144.0	0.043	0.941	8.161
HCHO (mg/m^3)	0.0	220.0	0.040	0.987	11.242
CO (ppm)	0.0	46.0	0.0	0.528	3.545
$PM_{1.0}$ ($\mu g/m^3$)	1.0	776.0	51.0	56.791	32.659
$PM_{2.5}$ ($\mu g/m^3$)	0.0	403.0	82.0	93.905	53.547
PM_{10} ($\mu g/m^3$)	0.024	400.0	75.0	100.402	113.554
$PM_{0.3}$ ($\mu g/m^3$)	1.0	288.0	68.0	74.852	41.001
AQI	2	298.0	122.0	119.94	56.581

Correlation matrix between pollutants and AQI categories:
The correlation matrix determine the relationship between each pollutant and the AQI

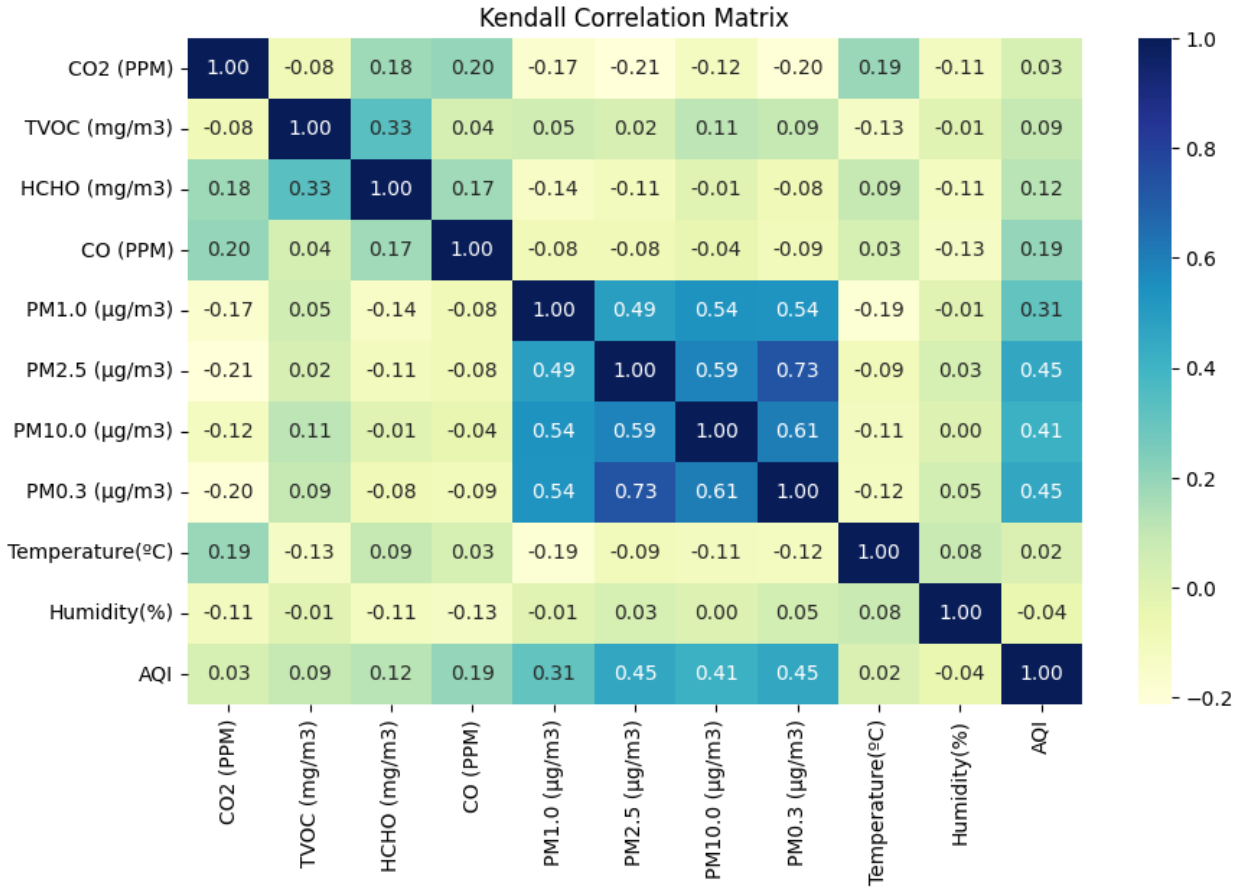


Figure 3.16: Kendall Correlation Matrix.

Kendall Correlation Matrix (3.16) shows Kendall's tau correlation coefficients between multiple pairs of variables, showing the strength and direction of monotonic

association for each pair. $PM_{2.5}$ (0.45), PM_{10} (0.41), $PM_{0.3}$ (0.45), and PM_1 (0.31) show strong correlation. CO (0.19), TVOC (0.09), HCHO (0.12), and CO_2 (0.03) show less correlation.

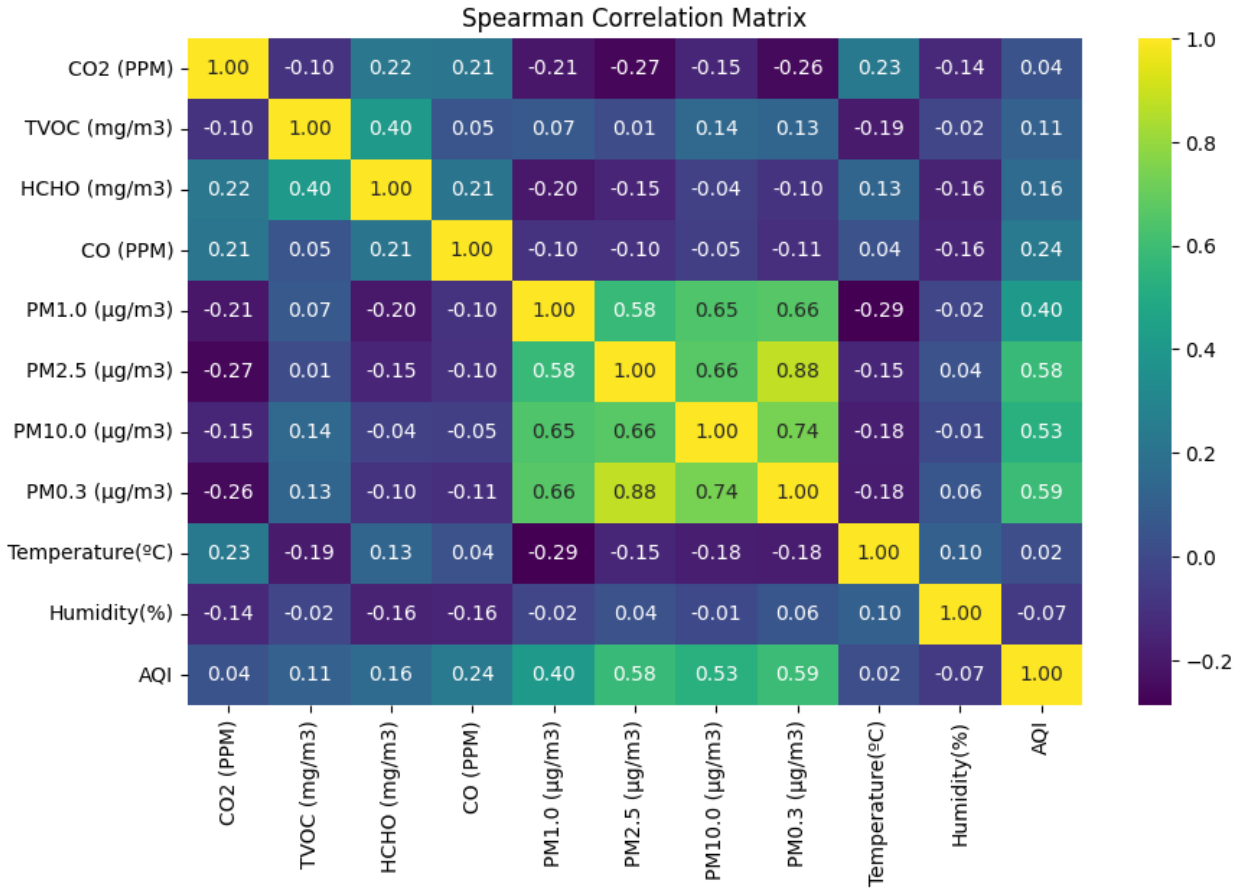


Figure 3.17: Spearman Correlation Matrix.

Spearman Correlation Matrix
(3.17) shows each cell Spearman's rho (ρ) value for a pair of variables. $PM_{0.3}$ (0.59), $PM_{2.5}$ (0.58), PM_{10} (0.53), and PM_1 (0.40) show strong correlation; CO (0.24), TVOC (0.11), HCHO (0.16), and CO_2 (0.04) show less correlation.

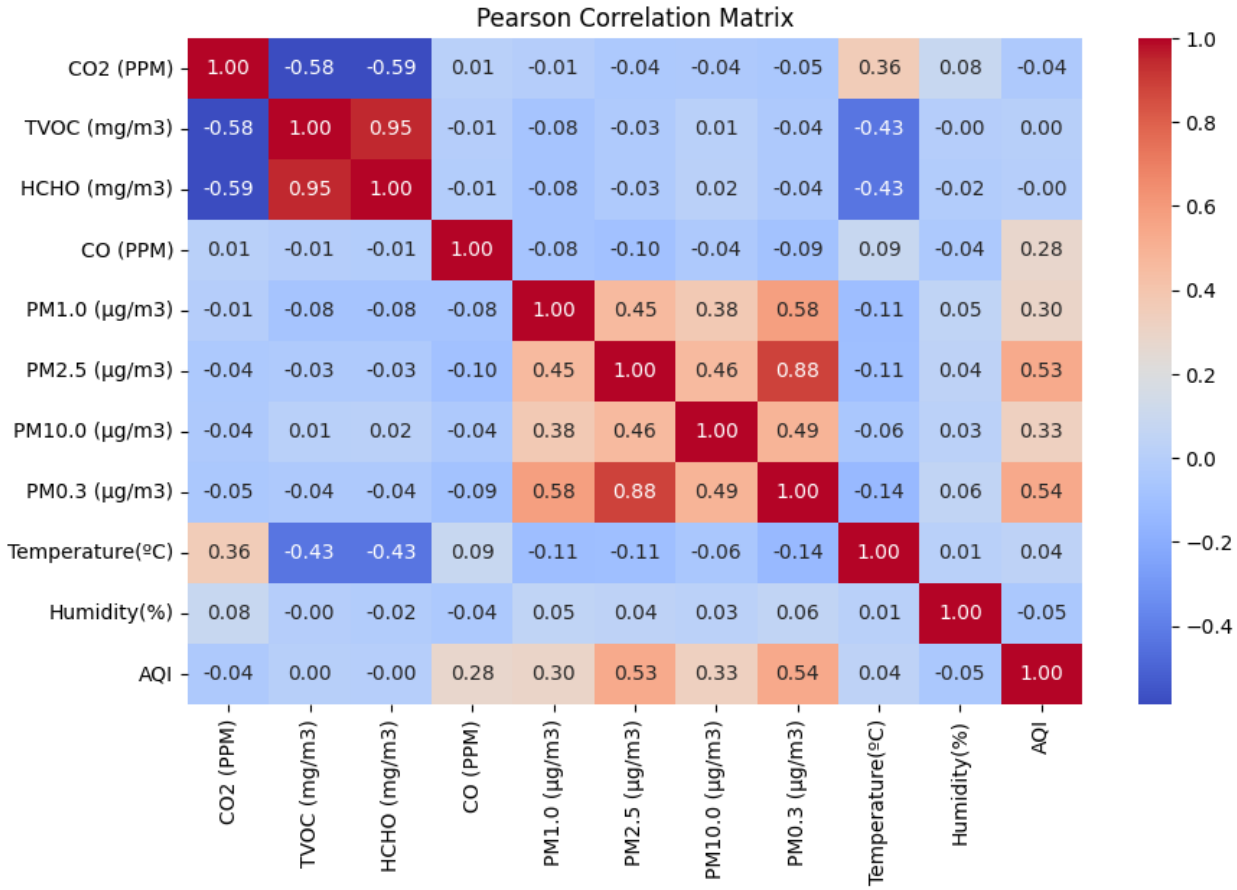


Figure 3.18: Pearson Correlation Matrix.

Pearson Correlation Matrix (3.18) shows linear relationships, which specifically measure how strongly two continuous variables are related in a straight-line pattern. $PM_{0.3}$ (0.54), $PM_{2.5}$ (0.53), PM_{10} (0.33), and PM_1 (0.30) show strong correlation, while CO (0.28) shows moderate correlation. TVOC (0) and HCHO (0) show no correlation, and CO_2 (-0.04) shows a slight negative correlation.
The key observation is that particulate matter is more responsible for the worst AQI than gaseous pollutants.

Class-wise analysis:
In this study, we follow six classes of AQI classification, which are standardized by the US EPA. Six classes are Good (0–50), Moderate (51–100), Caution (101–150), Unhealthy (151–200), Very Unhealthy (201–300), and Hazardous (301–500+).

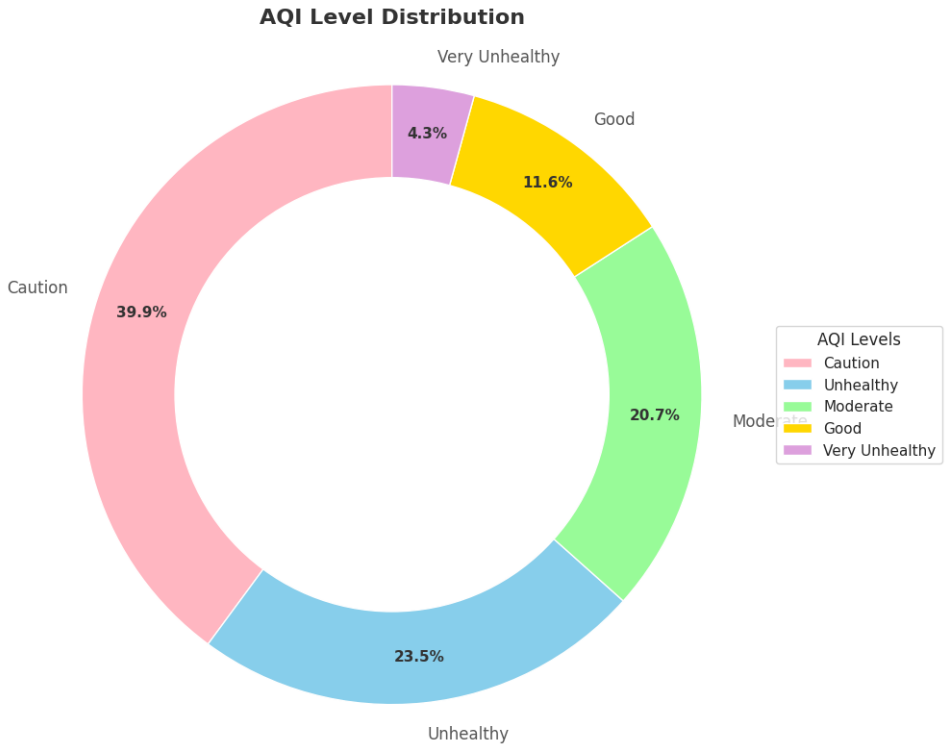


Figure 3.19: AQI level doughnut chart.

The donut chart (3.19) shows the caution class has the highest data at 39.9% (n=653); the second highest class is unhealthy at 23.5% (n=385); the moderate class has 20.7% (n=339); the good class has 11.6% (n=190); and the very unhealthy class has 4.3% (n=70) of the data. Additionally, no hazardous values were found anywhere.

Month-wise analysis:

AQI analysis: This box plot (3.20) shows monthly Air Quality Index (AQI) distribution. Which reveals a severe seasonal pattern, with persistently poor air quality from January to March and a significant peak of hazardous conditions in June. Notably, March was uniquely characterized by frequent, extreme pollution episodes, while June suffered from the highest sustained levels of pollution.

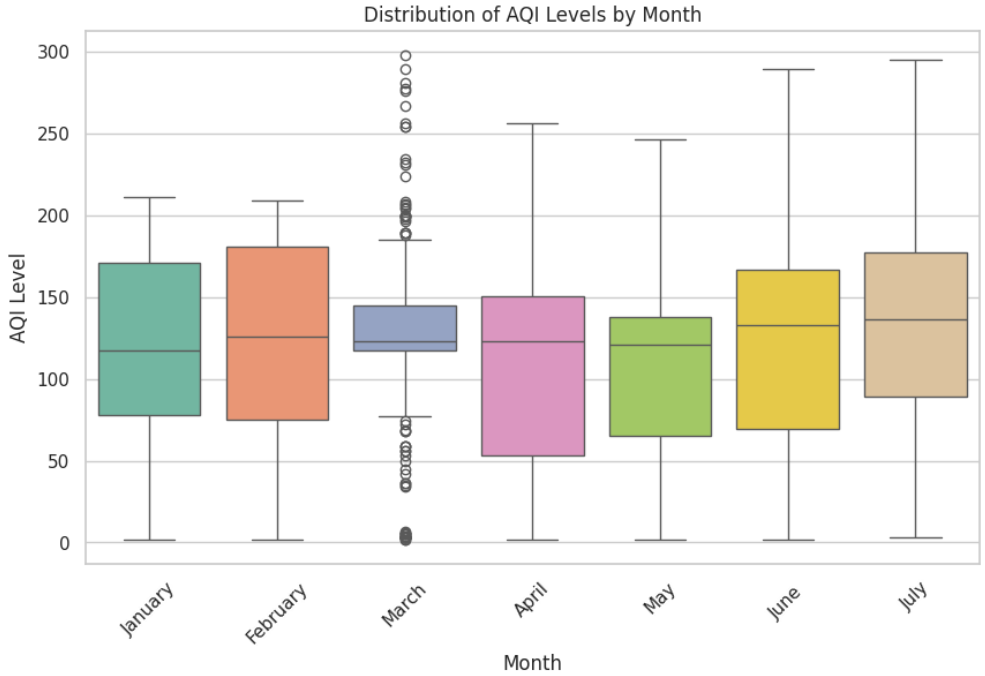


Figure 3.20: AQI distribution over months.

Gaseous Pollutants:

CO_2 and CO: The density of CO_2 (Fig. 3.22) and CO (Fig. 3.21) remains relatively stable across the observed months (January to July), with tight distributions and low variability. Median CO_2 levels show a slight increase in May, whereas CO levels are consistently low, with medians near the lower detection limit.

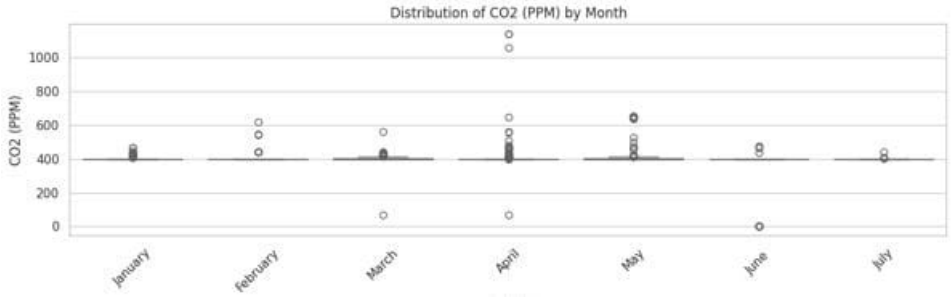


Figure 3.21: CO_2 over months.

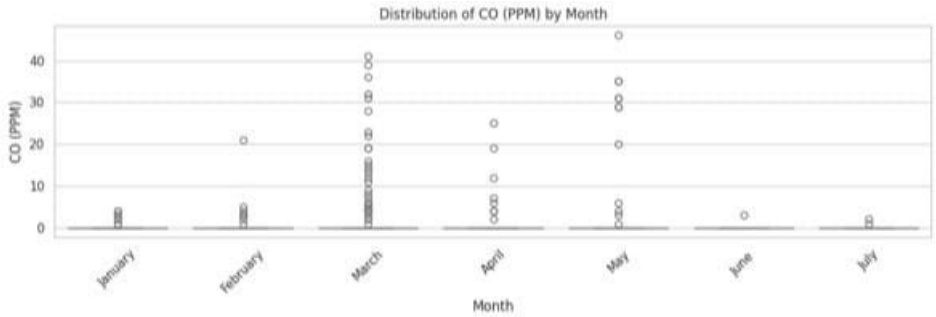


Figure 3.22: CO over months.

TVOC and HCHO: These elements show extremely low and stable concentrations from January to May and in July. However, the month of June is marked by a significant number of high-concentration outliers for both TVOC (Fig. 3.24) and HCHO (Fig. 3.23), suggesting the occurrence of episodic emission events during that period.

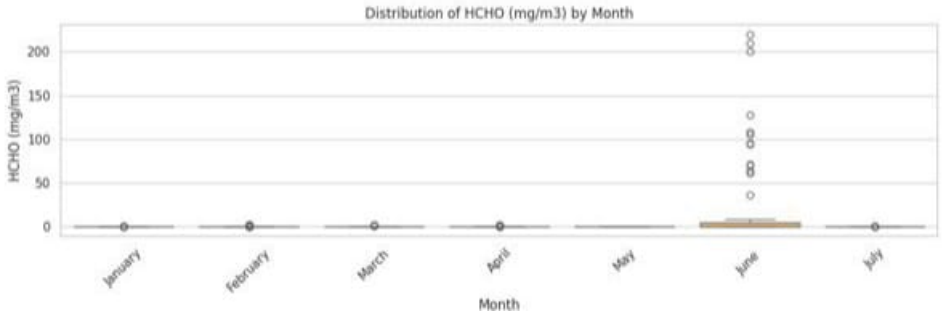


Figure 3.23: HCHO over months.

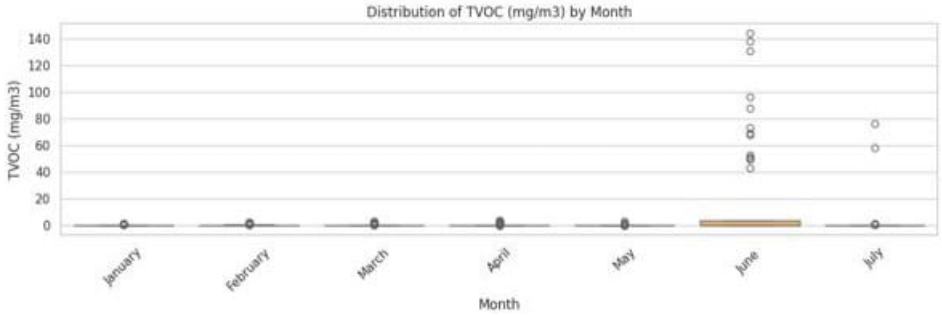


Figure 3.24: TVOC over months.

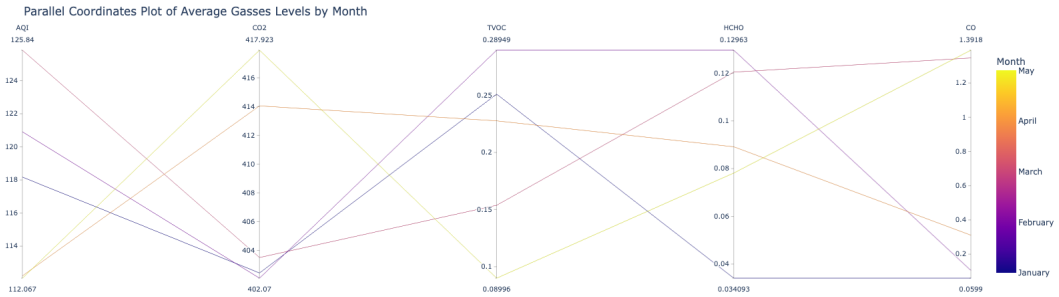


Figure 3.25: Parallel Coordinates Plot of Average Gas Levels by Months.

In Fig. 3.25, it shows there are complicated, inverse patterns of gaseous pollutants where higher levels of CO_2 and CO from winter to spring occur in exact opposition to a significant and simultaneous decrease in TVOC and HCHO concentrations.

Particulate Matters (PM):
All measured PM fractions, $PM_{0.3}$ Fig. 3.26, PM_{10} Fig. 3.27, $PM_{2.5}$ Fig. 3.28, and PM_1 Fig. 3.29, show a consistent and clear seasonal trend.

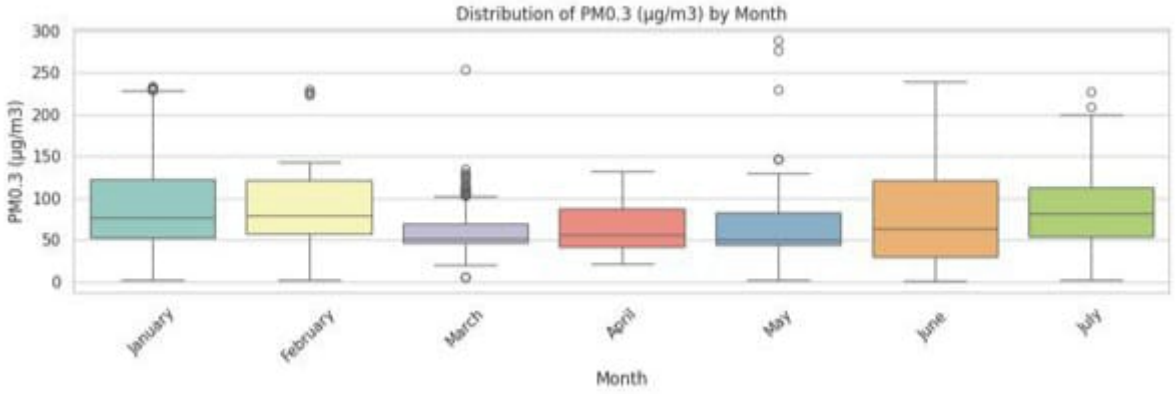


Figure 3.26: $PM_{0.3}$ over months

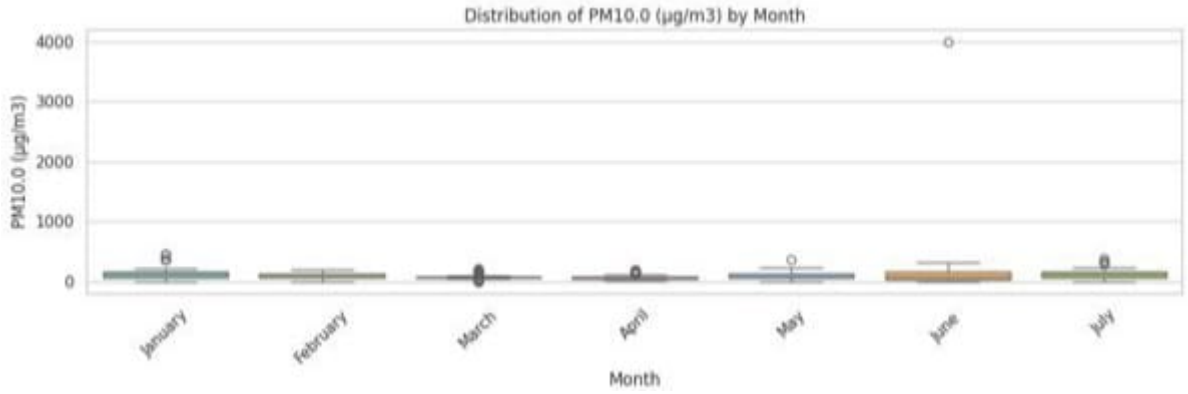


Figure 3.27: PM_{10} over months

Concentrations are highest in January, gradually decrease to their lowest levels around April, and then rise again to a secondary peak in June before declining in July. The variation, as indicated by the IQR, is most noticeable during the peak months of January and June.

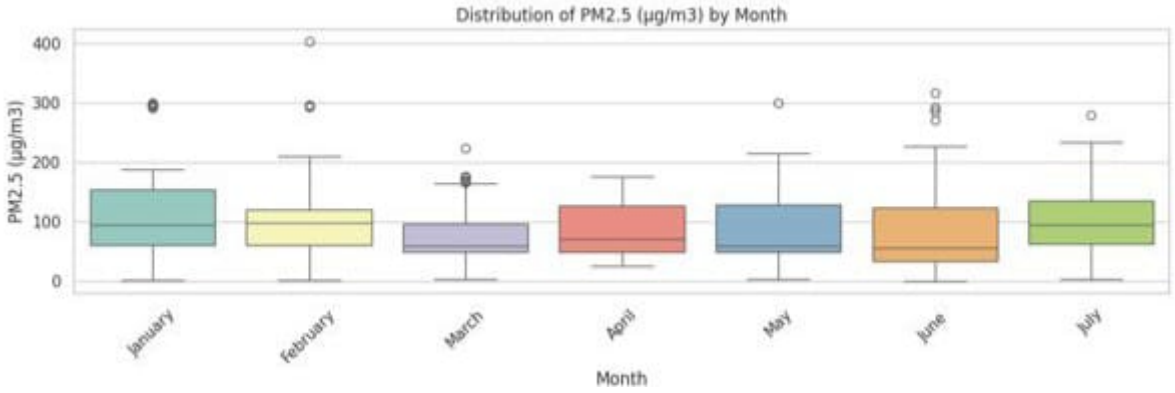


Figure 3.28: $PM_{2.5}$ over months

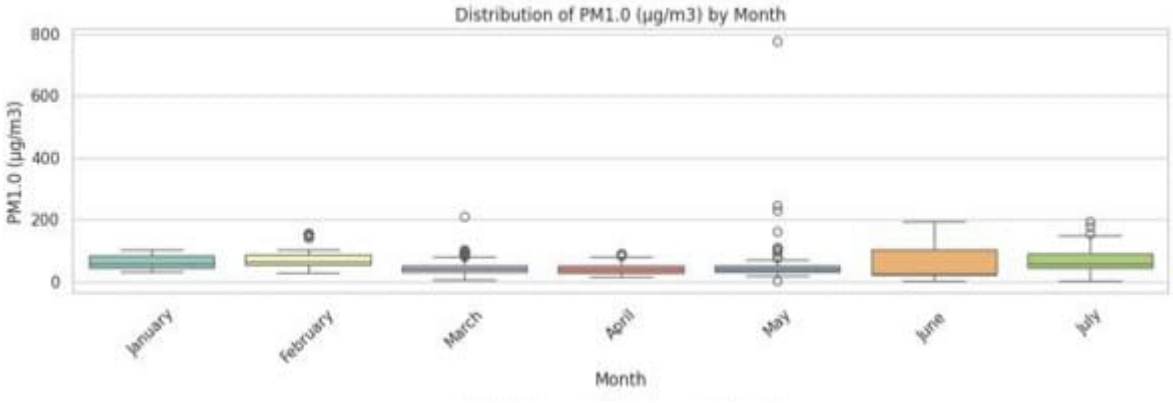


Figure 3.29: PM_1 over months

Fig. 3.30 shows a parallel coordinates plot of average PM levels by months. All particulate matter fractions are strongly correlated with each other and show a distinct seasonality with winter peaks in pollution (January-February) and a more significant, AQI-driven peak during the summer (June-July).

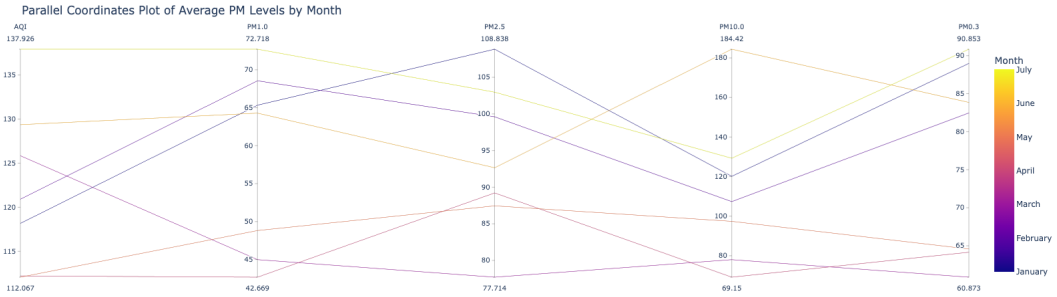


Figure 3.30: Parallel Coordinates Plot of Average PM Levels by Months.

Fig. 3.31 shows the monthly data reveal two contrasting patterns of pollution a typical, seasonal cycle for all fractions of particulate matter (PM) with summer and winter maxima, in contrast with overall gaseous pollutant stability interrupted by large, notable emissions for HCHO and TVOC during June. While the gaseous pollutants were comparatively stable, the particulate matter concentrations showed a distinct seasonal trend.

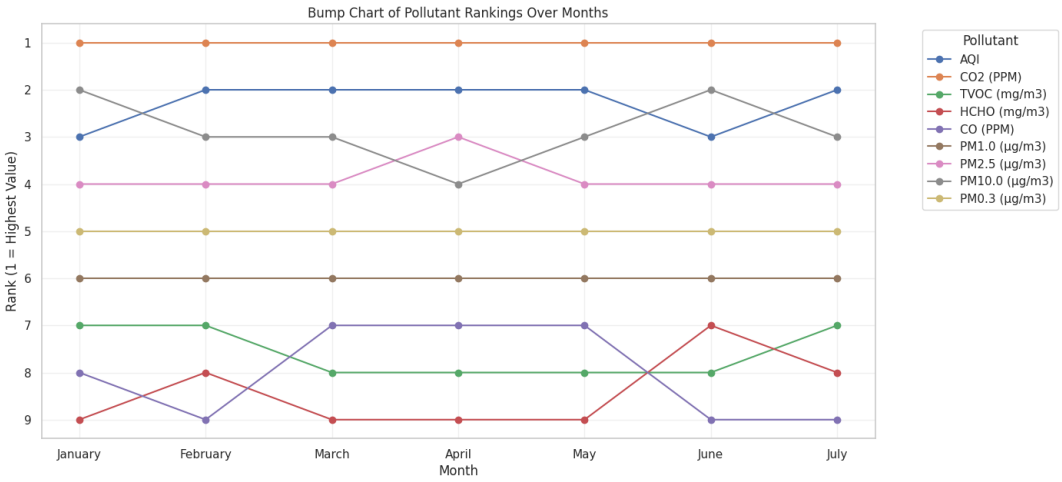


Figure 3.31: Bump Chart of all Pollutants and AQI over months

Location-wise analysis:

Datacount: The bar chart Fig. 3.32 presents a bar chart detailing the distribution of data points across 19 distinct locations. The data collection appears relatively balanced for most sites, with the number of data points per location generally ranging from 75 to 100. The 'Fire Exit (6th floor)' provided the maximum number of samples (n=110), closely followed by 'Inside Kitchen' (n=107) and 'Pond' (n=105).

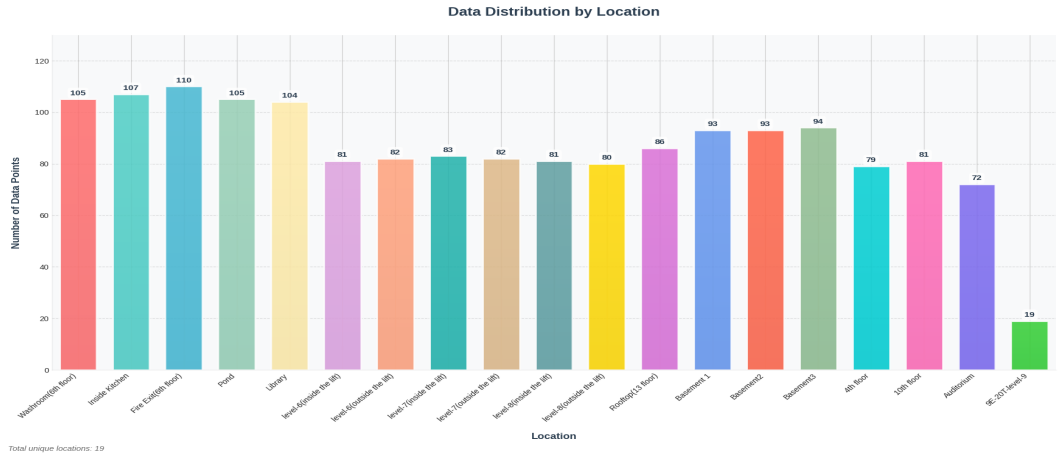


Figure 3.32: Location-wise data count.

On the other hand, the '9E-207-level-9' location is a significant outlier, contributing the lowest count of only 19 data points, indicating its underrepresentation in the dataset compared to the other locations.

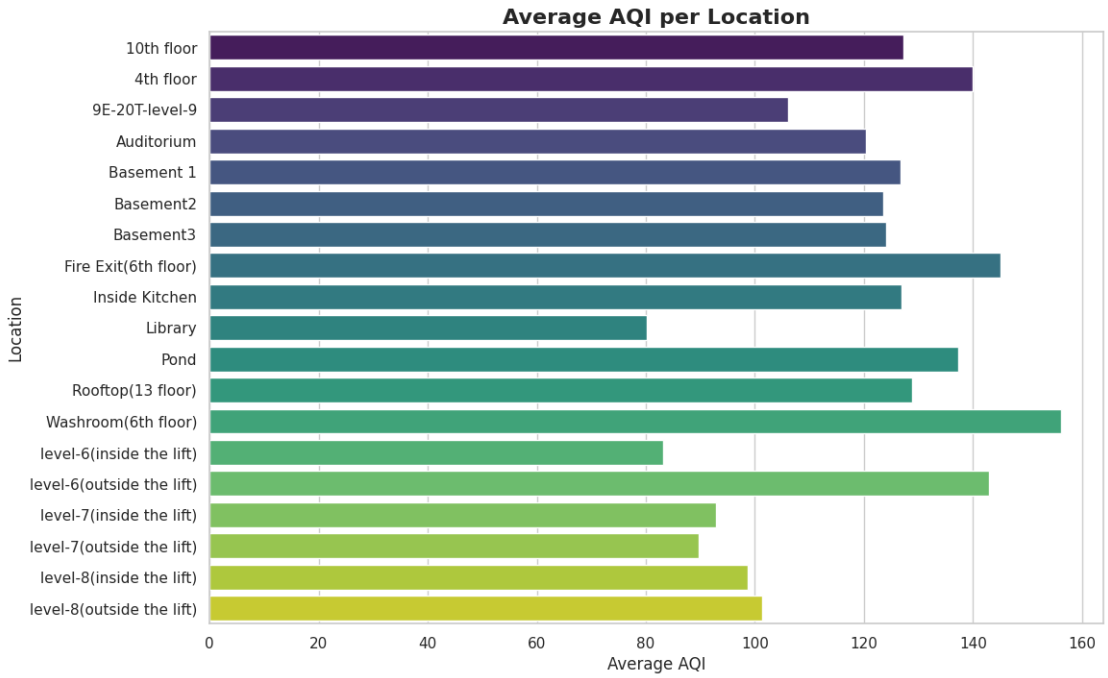


Figure 3.33: Horizontal Bar chart of average AQI over locations

AQI analysis:
The horizontal bar chart Fig. 3.33 shows the highest AQI in the washroom, and fire exit is approximately 145 to 155. The library shows the lowest AQI and other areas are in a stable pattern.

Gaseous Pollutants:
A comparative analysis of gaseous pollutants Fig. 3.34 shows distinct concentration patterns across the various locations, as visualized on a logarithmic scale to accommodate values spanning several orders of magnitude. Therefore, CO_2 concentrations remain stable at approximately 400 PPM, consistent with ambient atmospheric levels, while CO is negligible, mostly in all areas except some basement areas, suggesting an absence of significant combustion sources.

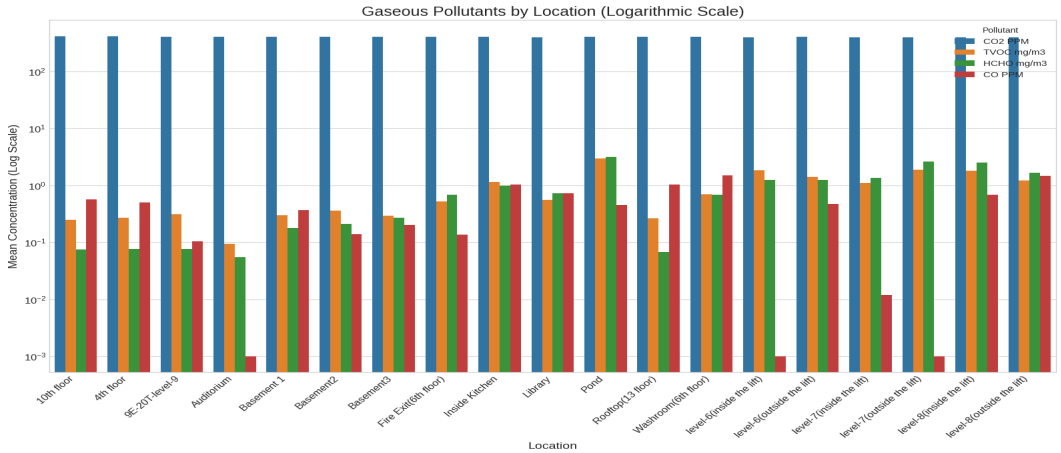


Figure 3.34: All gas concentrations over location (bar chart)

In stark contrast, TVOC and HCHO exhibit significant spatial variation, with

markedly elevated levels in enclosed environments such as the kitchen, washroom, and basement areas. This suggests that localized emission sources, likely related to specific activities or materials within these areas, are the primary drivers of indoor air quality degradation, whereas background gas levels are largely uniform.

Particulate Matter (PM):
The spatial distribution of particulate matters Fig. 3.35 shows significant variability across the different monitored locations. PM mass concentrations are highly dependent on the specific environment, with the highest levels of PM_{10}, $PM_{2.5}$, PM_1, and $PM_{0.3}$ consistently recorded in indoor areas such as the 'Library', 'Basement3', and the '4th' and '10th' floors. Conversely, the 'Pond' and 'Fire Exit' locations exhibited the lowest PM levels, likely due to natural ventilation and lower human activity. In areas with high pollution, the concentrations of all particle sizes trend together, indicating common emission or ventilation sources. Notably, $PM_{2.5}$ constitutes a substantial fraction of the total PM_{10} mass in these locations, highlighting a potential health concern associated with exposure in these specific indoor environments.

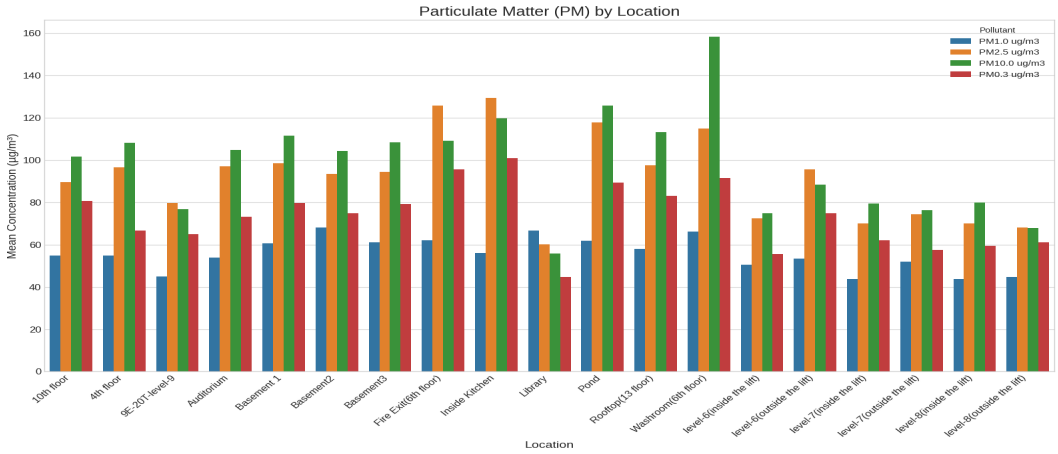


Figure 3.35: All PM's concentration by location (Barchart).

Indoor vs. Outdoor Analysis : Fig. 3.36

Indoor analysis:
The indoor air quality analysis indicates robust spatial heterogeneity, driven by extremely localized emission sources. The indoor locations like the Washroom (6th floor), Inside Kitchen, Fire Exit (6th floor), all three Basements, 4th and 10th floors, and Auditorium are found to have poor air quality. In particular, the Washroom has the highest overall AQI, the Kitchen is a significant source of TVOCs, and the Basements are hot spots for CO and HCHO, likely from car exhaust. By sharp contrast, other indoor environments like the library and the areas in the lifts show consistently low pollutant levels. This comparison emphasizes the fact that despite many of the indoor areas suffering from heavy internal pollution, effective local air cleaning can efficiently create clean air areas within the building.

Outdoor analysis:
Outdoor location analysis implies that ambient air quality in the immediate vicinity of the building is not good. Places like Pond, Rooftop (13th floor), and open spaces

on floors 6, 7, and 8 always record high AQI values. Both gaseous pollutants (TVOC, HCHO, and CO) and all fractions of particulate matter have high concentrations as per the data. This extremely polluted external environment is a constant external hazard, a steady supply of contaminants that percolate into the building and must be addressed by its ventilation and filtering equipment.

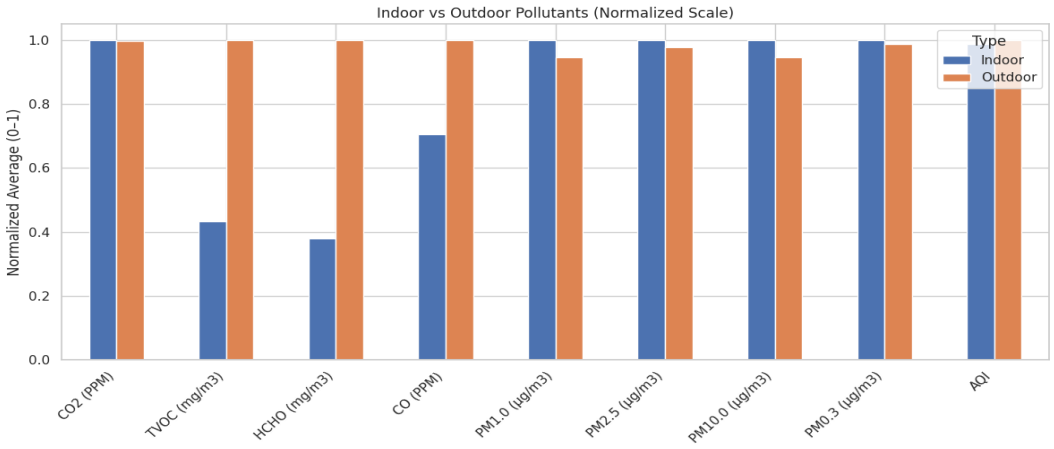


Figure 3.36: Indoor vs Outdoor (Bar chart).

To sum up, outdoor air is consistently heavily polluted; indoor air quality varies dramatically due to severe, localized internal emission hotspots.

3.4 Machine Learning Models

3.4.1 CatBoost (Categorical Boosting)

Ensemble of Decision Trees (Gradient Boosting):

- CatBoost creates a group of decision trees, where each tree learns from the mistakes made by the previous ones before it makes predictions.
- Prediction equation:

Fig. 3.37 illustrates the conceptual workflow of the CatBoost model.
The prediction function in CatBoost can be represented as (Eq. 3.1):

$$F(x) = F_0(x) + \sum_{m=1}^{M} f_m(x) \tag{3.1}$$

where,
$F(x)$: Final prediction for input x,
$F_0(x)$: Initial prediction,
$f_m(x)$: Output of the m-th tree, and
M : Total number of trees.

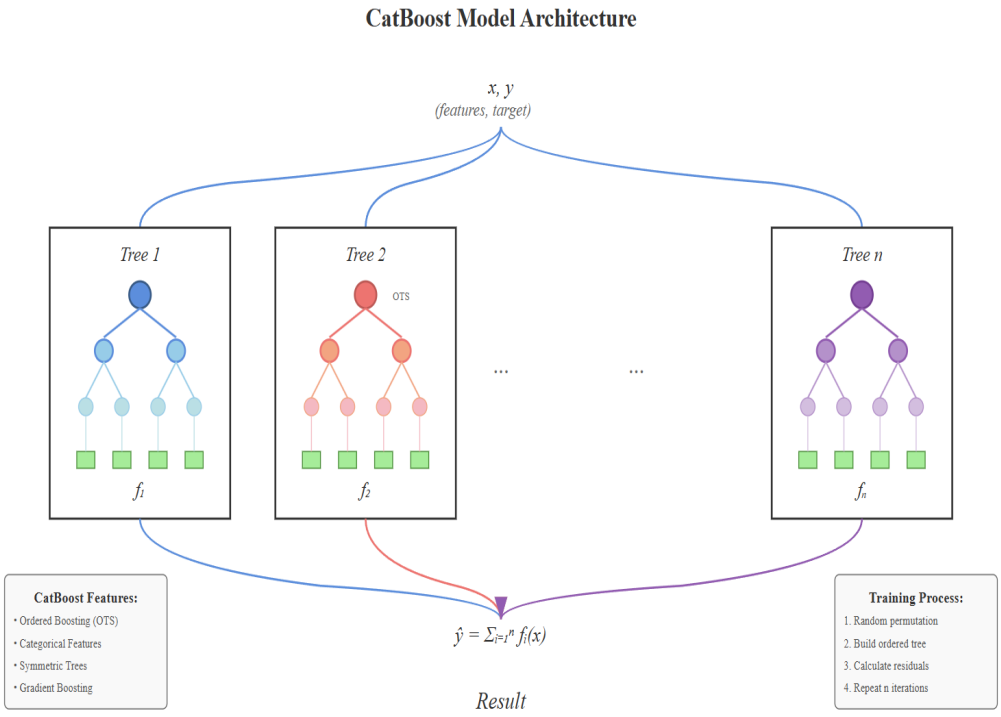


Figure 3.37: CatBoost Model Architecture.

Ordered Boosting & Symmetric Trees:

- Ordered Boosting: CatBoost uses a unique method called ordered Boosting to train the model more securely. The model does not accidentally 'peek' at the answers while learning because it shuffles the training data in a smart way. This way helps the model to learn genuinely, avoid overfitting data and makes better predictions.
- Symmetric Trees: At the same level all the nodes of the tree split the data in the same way, which makes the tree balanced and helps it to work faster which gives better performance.

Loss Function (Objective):

- For multiclass classification, CatBoost minimizes multiclass log loss:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{3.2}$$

where,
y_i: True label of the i-th sample, and
$\hat{y}_i$: Predicted probability for class i.

- The model looks at how wrong the current predictions are — meaning the difference between actual and predicted values — and it learns from the errors to make better predictions in the upcoming iterations.

Gradient and Hessian

- CatBoost uses two types of information. One is how much the error is changing, which means the gradient, and the other one is how fast it's changing, which is called the Hessian, to improve each tree.

Regularization and Early Stopping

- Regularization: This adds a small penalty to every model to stop them from learning noise in the data.
- Early Stopping: It means the model stops training early if the accuracy or loss stops improving after several rounds, avoiding it from overtraining.

Prediction and Evaluation

- The final model is the sum of all trees' outputs, producing a probability distribution over AQI categories.

3.4.2 XGBoost (Extreme Gradient Boosting)

Ensemble of Decision Trees (Gradient Boosting Framework)

- **Core Principle:** XGBoost works the same as CatBoost, where each tree learns from the mistakes made by the previous ones before it makes predictions.

Figure 3.38 shows the structure of the XGBoost model.
The prediction for the i-th instance can be expressed as (Eq. 3.3):

$$\hat{y}^{(i)} = F_0(x_i) + \sum_{m=1}^{M} f_m(x_i) \tag{3.3}$$

where,
$\hat{y}^{(i)}$: Final prediction for input,
$F_0(x_i)$: Initial prediction,
$f_m(x_i)$: Output of the m-th tree, and
M : Total number of trees.

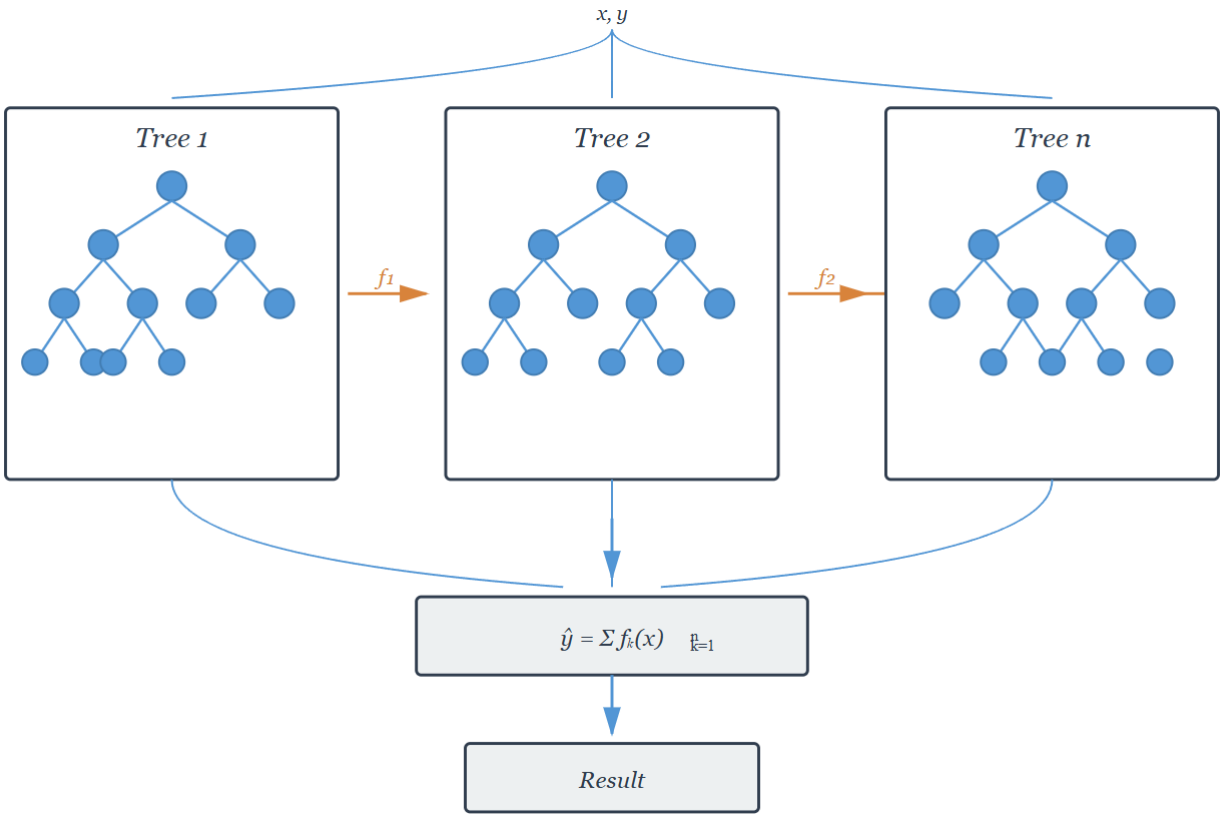


Figure 3.38: XGBoost Model Architecture.

Gradient Boosting Optimization in XGBoost

Loss Function: For classification, XGBoost minimizes multiclass log-loss given by (Eq. 3.4):

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{3.4}$$

where,
y_i : True label for the i-th instance.
$\hat{y}_i$: Predicted probability for class i.
Gradient and Residuals: At each iteration, XGBoost computes the gradient — the difference between the model's current predictions and the actual values — to determine how to improve the next tree (Eq. 3.5):

$$r_i^{(t-1)} = -\frac{\partial \hat{y}_i^{(t-1)}}{\partial L(y_i, \hat{y}_i^{(t-1)})} \tag{3.5}$$

Each new tree is trained to correct these residuals, updating the model as (Eq. 3.6):

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \tag{3.6}$$

where η is the **learning rate** controlling each tree's contribution.
Regularization: XGBoost discourages trees which are too complex to handle, so the model does not memorize the training data and perform better with new data.
Prediction and Evaluation: The model summing the outputs of all trees and uses that to predict AQI category for each of the samples in the test set.

3.4.3 Random Forest

Multiple independent decision trees are constructed by Random Forest and their outputs are combined for final prediction. (Figure 3.39).
Core Principle: Random Forest first builds many decision trees and then combines all the tree results to make actual predictions.
Bootstrap Aggregation (Bagging): Each tree is trained on a random sample of the training data is called bootstrapping.
Prediction Aggregation: For the classification part, each of the trees votes for a class and the most voted class becomes the final prediction (Eq. 3.7):

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_N(x)\} \tag{3.7}$$

where,
$T_i(x)$ is the prediction of the i-th tree, and
N is the total number of trees.

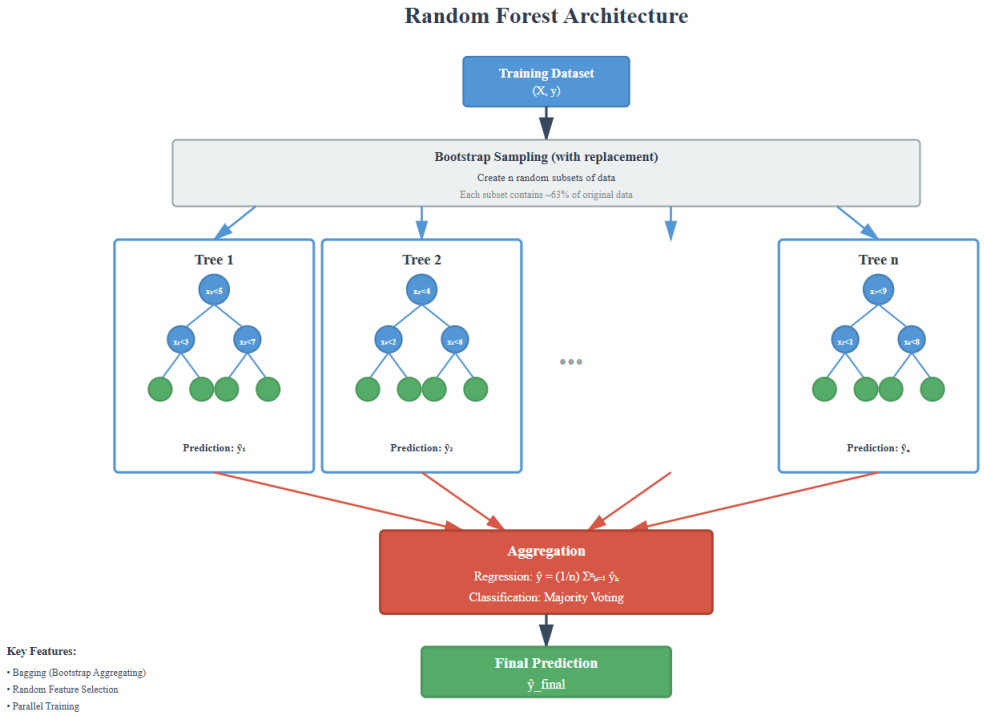


Figure 3.39: Random Forest Model Architecture.

Decision Tree Splitting: Each of the trees divides the data using the feature values to make classes as separate as possible and chooses the best split which gives the accurate information.
Gini Impurity:

$$G = 1 - \sum_{k=1}^{K} p_k^2 \tag{3.8}$$

Entropy:

$$H = -\sum_{k=1}^{K} p_k \log(p_k) \tag{3.9}$$

In (Eq. 3.8), (Eq. 3.9) where p_k is the proportion of samples belonging to class k within the node.
Random Feature Selection: At each split, the tree only looks for the random selection of the features, which makes the tree different from each other.

3.4.4 LightGBM (Light Gradient Boosting Machine)

Ensemble Gradient Boosting Framework, LightGBM builds a strong model by adding simple trees step by step. Each tree learns and fixes their mistakes by the help of previous trees. (Figure 3.40).
Prediction Equation:

$$\hat{y}^{(i)} = F_0(x_i) + \sum_{m=1}^{M} f_m(x_i) \tag{3.10}$$

In (Eq. 3.10) where,
$F_0(x_i)$ is the initial prediction,
$f_m(x_i)$ is the output of the m-th tree, and
M is the total number of trees.

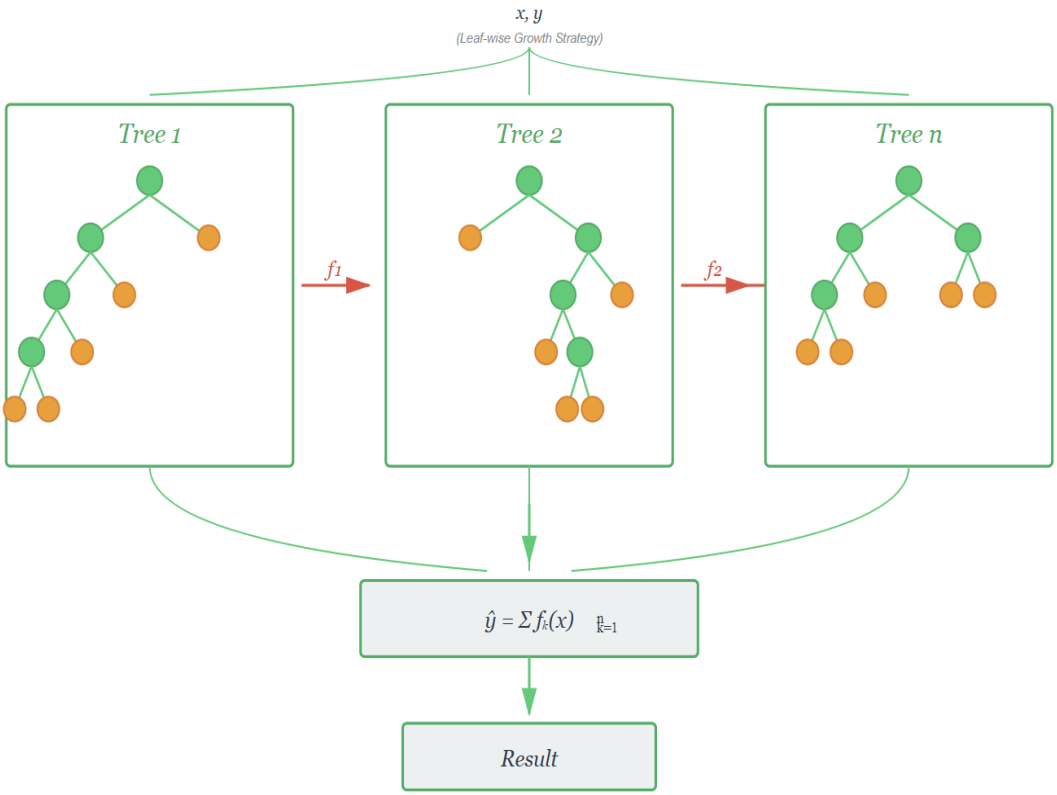


Figure 3.40: LightGBM Model Architecture.

Leaf-wise Tree Growth: Unlike most boosting methods that grow trees according to their levels but in LightGBM it grows trees leaf by leaf.
Histogram-based Splitting: LightGBM groups the feature values into bins and uses those bins to decide the best split. This makes the tree use less memory and faster to build.

Loss Function: For classification, LightGBM minimizes the multiclass log-loss (cross-entropy)(Eq. 3.11):

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{3.11}$$

Here, y_i is the true label
$\hat{y}_i$ is the predicted probability for class i.

Gradient-Based One-Side Sampling (GOSS)

- During the training period, LightGBM keeps the data which are hard to predict and randomly picks some of the easy ones. In this way, the model learns from the most important and challenging examples.

Exclusive Feature Bundling (EFB):

- LightGBM combines the features which don't appear together to make the data simpler which helps the model to train faster and especially helps when there are multiple features and mostly empty values.

Regularization and Efficiency:

- Using regularization helps the model avoid memorizing the training data. Also train multiple parts at the same time for making it faster and able to handle big datasets.

3.4.5 LSTM (Long Short-Term Memory Network)

LSTM is a recurrent neural network (RNN) variant that captures both short-term and long-term dependencies in sequential data such as pollutant readings (Figure 3.41).
LSTM Cell Structure: Each LSTM unit consists of a memory cell and three gates:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad \text{(Forget Gate)} \tag{3.12}$$
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad \text{(Input Gate)} \tag{3.13}$$
$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \quad \text{(Cell Candidate)} \tag{3.14}$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad \text{(Cell State Update)} \tag{3.15}$$
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad \text{(Output Gate)} \tag{3.16}$$
$$h_t = o_t \odot \tanh(C_t) \quad \text{(Hidden State Output)} \tag{3.17}$$

where,
x_t is the input at time t,
h_{t-1} the previous hidden state,
σ the sigmoid activation, and
$\odot$ denotes element-wise multiplication.
Dropout Layers: It randomly turns off some neurons while training so that the model does not rely on any part which helps it to learn better and perform well on new data.
Dense Layers:

- **Hidden Dense Layer:** Applies a ReLU activation, $ReLU(z) = \max(0, z)$.
- **Output Dense Layer:** Uses the softmax function to generate class probabilities across AQI categories.

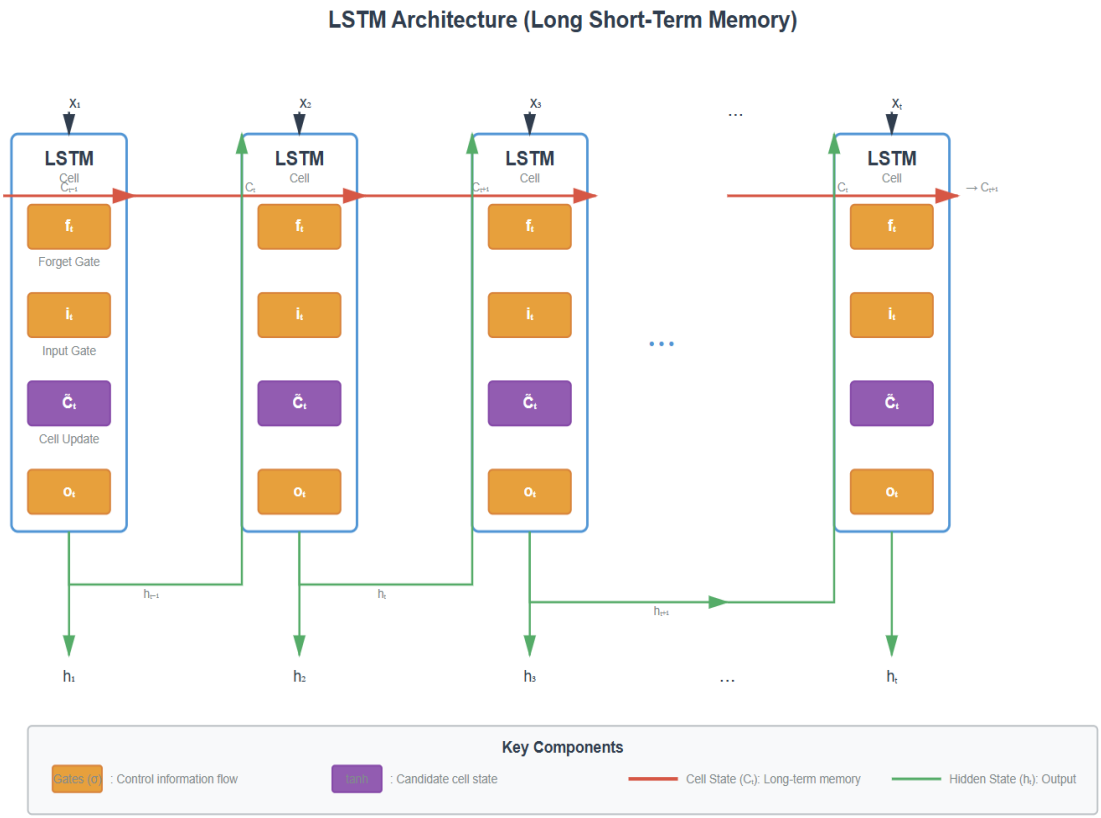


Figure 3.41: LSTM Model Architecture.

Loss Function: Categorical Cross-Entropy (Eq. 3.18):

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{3.18}$$

Optimizer: Adam optimizer adapts learning rates for faster convergence.
Prediction and Evaluation: Predicts the AQI category with the highest probability. Performance is evaluated using accuracy, precision, recall, F1-score, and a classification report.

3.4.6 Support Vector Machine (SVM)

SVM is a supervised learning method that finds the next hyperplane to separate the data into different classes, even though there are many features.(Figure 3.42).

Kernel Trick: The RBF kernel changes the data into the higher dimensional space, which helps SVM to handle the non linear data and complex patterns.
The kernel function is:

$$K(x, x') = \exp(-\gamma \|x - x'\|^2) \tag{3.19}$$

In (Eq. 3.19), where γ controls the influence of each training instance.

Decision Function: The SVM predicts the class label by computing:

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\right) \tag{3.20}$$

In (Eq. 3.20) where,
α_i are Lagrange multipliers,
y_i class labels,
$K(x_i, x)$ the kernel function, and
b the bias term.

Optimization Objective: Maximize the Margin: SVM tries to make the gap as wide as possible because of the diving line and closest points of different classes. Soft Margin: The regularization parameter C. balances making the margin wide and keeping classification error small

Probability Estimates: With `probability=True`, SVM uses Platt scaling to convert decision function outputs into probability estimates for each class.

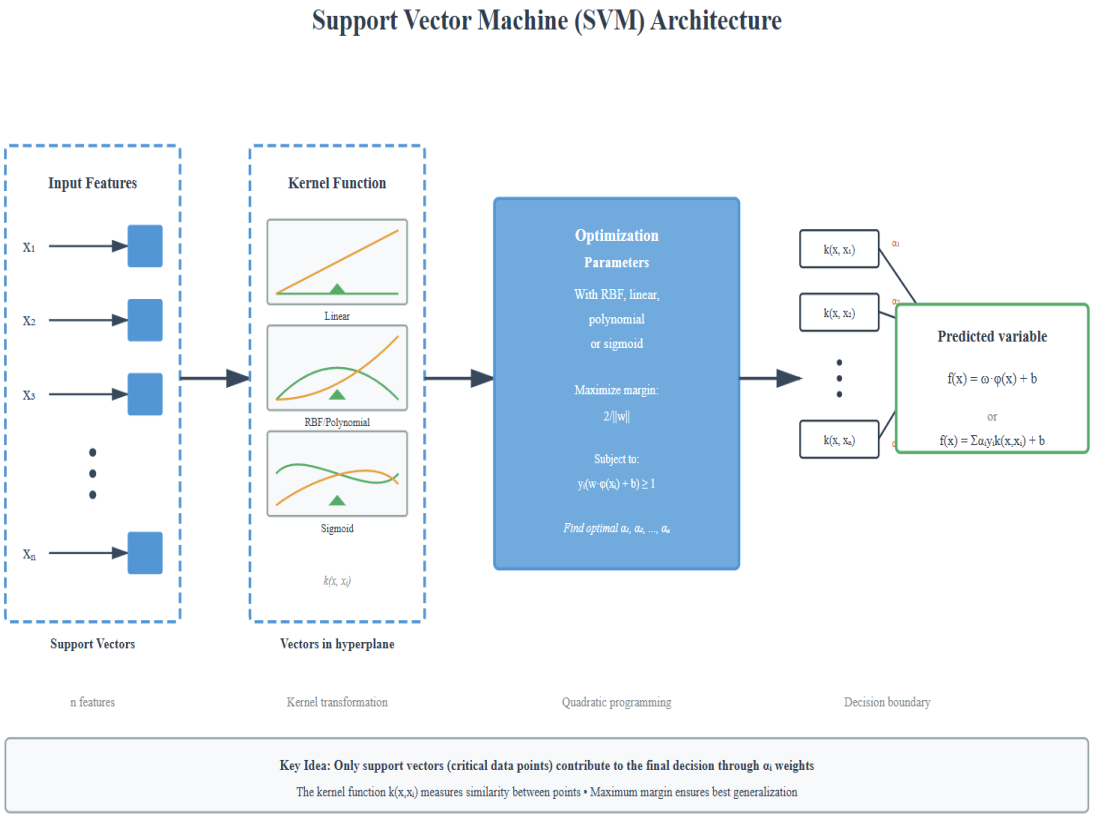


Figure 3.42: SVM Model Architecture.

Prediction and Evaluation:

Prediction: SVM predicts AQI categories for the test data by choosing the class which got the highest score for each sample data.
Evaluation: Model performance is checked using accuracy, precision, recall, F1-score, and a classification report tells us how well it predicts.

3.4.7 GRU (Gated Recurrent Unit)

The GRU layer also processes the sequential data and can remember important past information in pollutant features. The model understands the patterns over time.(Figure 3.43).
GRU Cell Equations:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad \text{(Update Gate)} \tag{3.21}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad \text{(Reset Gate)} \tag{3.22}$$

$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \quad \text{(Candidate Activation)} \tag{3.23}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad \text{(Hidden State Update)} \tag{3.24}$$

Here,
x_t is the input at time t,
h_{t-1} is the previous hidden state,
σ is the sigmoid activation,
$\circ$ denotes element-wise multiplication.

Dropout Layer: Randomly drops a fraction of units during training to prevent overfitting and improve generalization.

Dense Output Layer: It takes GRU outputs to AQI category probabilities. Uses softmax function for handling multiple categories. For all AQI categories it produces a probability.
Loss Function: Categorical Cross-Entropy:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{3.25}$$

In (Eq. 3.25) where,
y_i: True label, and
$\hat{y}_i$: Predicted probability for class i
Optimizer: It is an optimizer which automatically adjusts the learning speed. Also uses the errors for updating the models weight and helps the model to learn faster.
Prediction and Evaluation: The main trained models give a chance for each AQI category. The category with the highest chance is chosen as the prediction.

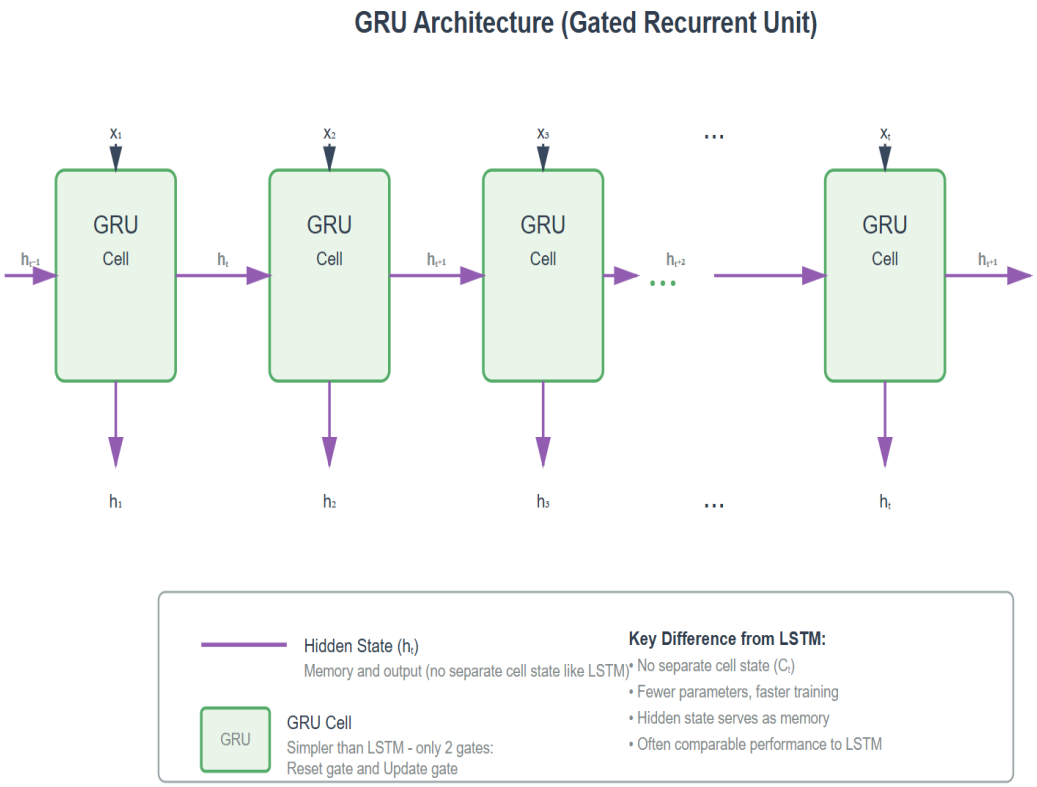


Figure 3.43: GRU Model Architecture.

3.4.8 Hybrid Model

Overview:
In Figure 3.44 shows the proposed model architecture. The proposed model is a tabular classification model with a hybrid architecture of deep learning and machine learning. The deep learning part is used for extracting feature embeddings from multivariate air pollutant data and a machine learning (random forest classifier) for refined classification. Additionally, the final prediction is observed through a weighted soft-voting ensemble, combining both components to increase robustness and improve categorical Air Quality Index (AQI) prediction.
Data flow pipeline:
Input features (CO_2, TVOC, HCHO, $PM_{0.3}$, $PM_{1.0}$, $PM_{2.5}$, and $PM_{10.0}$) are reshaped for Conv1D layer input.
The CNN block finds the local patterns of features and complex abstractions, which are followed by the Feature Attention Mechanism to capture the most informative features. Moreover, this full process creates a 128-dimensional embedding vector as the final feature presentation.
After that the embedding is directly captured with path A (Neural Network) which is a fully connected softmax classifier. Also, the embedding is used as input to the path B (Random Forest classifier) model to perform categorical classification. The probabilistic outputs of the neural network and random forest are combined through weighted soft voting, giving the random forest greater weight for its robustness with tabular embeddings. The final classification prediction is the highest combined probability.

Neural Network Architecture (Feature Extractor):
The CNN block is used as a backbone with feature attention to capture the spatial dependencies among pollutants (Figure 3.44).
Input layer:
The input shape is (7,1), where 7 is the pollutant feature and 1 is the channel dimension.
Input format,
$X = [x_1, x_2, \ldots, x_F],\ X \in \mathbf{R}^{(F \times 1)}$
where $F = 7$, the number of pollutant features.

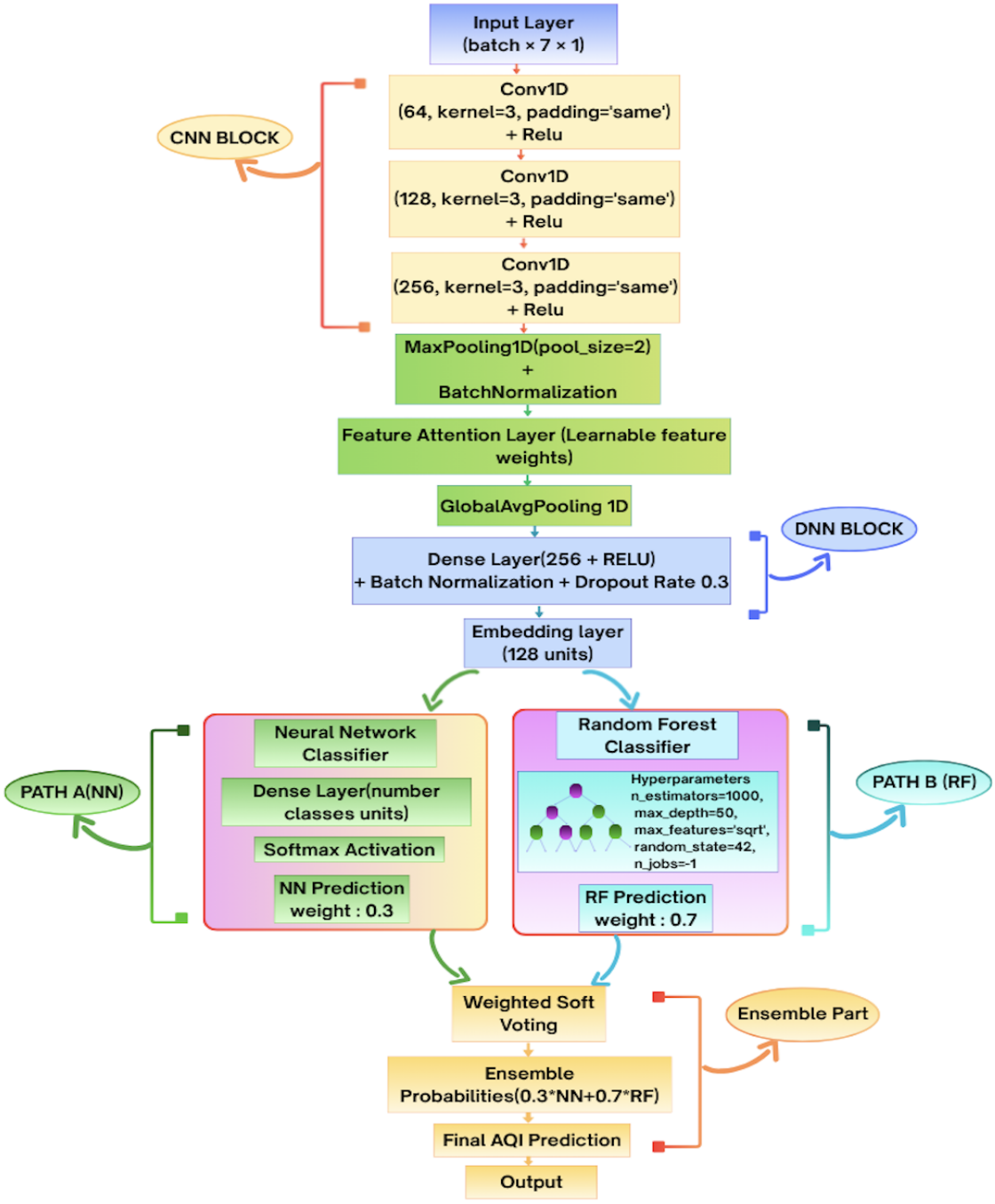


Figure 3.44: Hybrid Model Architecture.

Convolutional Layers:
Conv1D (64 filters, kernel size = 3) and ReLU activation, Conv1D (128 filters, kernel size = 3) and ReLU activation, Conv1D (256 filters, kernel size = 3) and ReLU activation.
Each Conv1D layer can be defined as,

$$h^{(l)} = ReLU(W^{(l)} * h^{(l-1)} + b^{(l)}) \tag{3.26}$$

In (Eq. 3.26), where
"*" = 1D convolution operation;
$W^{(l)}, b^{(l)}$ = trainable weights and biases of layer l;
$h^{(0)} = X$ is the input feature vector.

Pooling and Normalization:
MaxPooling1D with pool size = 2 decreases dimensionality while preserving major features. Batch normalization stabilizes training and accelerates convergence.

Feature Attention Layer:
A custom attention technique learns and allocates importance weights to each feature map. These attention scores are normalized using softmax and applied through multiplication, directing the model's concern to the most focused pollutants for classification.
The weighted feature representation becomes:

$$\alpha_i = \frac{\exp(x_i \cdot w)}{\sum_{j=1}^{F} \exp(x_j \cdot w)} \tag{3.27}$$

In (Eq. 3.27), where
x_i = feature representation of pollutant i;
w = trainable attention weight vector;
α_i = normalized attention score for feature i.
GlobalAveragePooling1D condenses each feature map into a global representation.

Dense layers:
Dense (256 filters) and ReLU activation with Batch Normalization and Dropout (0.3) concentrate representations and prevent overfitting. Dense (128 filters) with ReLU activation, which is named "embedding," produces the final 128-dimensional embedding vector, which acts as a compressed representation of air quality patterns.

Softmax Classification Head (NN path):
Dense (num_classes, softmax) outputs class probabilities for AQI levels.

Neural Network Classifier (Path A):
Softmax output for class c:

$$P_{NN}(c|X) = \frac{\exp(W_c^T e + b_c)}{\sum_{k=1}^{C} \exp(W_k^T e + b_k)} \tag{3.28}$$

In (Eq. 3.28), where C is the number of AQI categories.
Softmax Classification Head: Dense (number of classes, softmax), which predicts

probabilities for AQI levels. The loss function is categorical cross-entropy, and the optimizer is Adam with a learning rate of 0.0008.

Random Forest Classifier on Learned Embeddings (Path B):
The embeddings followed by the CNN are extracted and provided as input to a random forest classifier. Random Forest hyperparameters Number of trees (number of estimators) = 1000, Maximum depth = 50, Feature selection at splits = $\sqrt{(\text{number of features})}$, Parallel training with all available CPU cores.
The Random Forest classifier praises the deep neural network by building an ensemble of various decision trees over the embedding space. By combining randomness and aggregation, each tree captures unique partitions of the feature space, and their collective output forms a robust, accurate classifier. In the hybrid model, Random Forest converts the 128-dimensional neural embeddings into precise class boundaries, strengthening the deep feature extractor with a rule-based ensemble that markedly improves prediction accuracy and stability.
The RF predicts class probabilities by averaging across trees:

$$P_{RF}(c|e) = \frac{1}{T}\sum_{t=1}^{T} P_t(c|e) \tag{3.29}$$

In (Eq. 3.29), where:
T = number of trees = 1000;
$P_t(c|e)$ = probability assigned to class c by tree t.

Ensemble Strategy:
To explore the strengths of both models, a weighted soft-voting ensemble is applied (Eq. 3.30):

$$P_{ensemble}(c) = 0.7 \times P_{RF}(c) + 0.3 \times P_{NN}(c) \tag{3.30}$$

The greater weight is given to the Random Forest classifier to reflect its stronger ability to capture decision boundaries in the embedding space; moreover, the neural network provides complementary information through its direct softmax outputs.

Tuning:
The proposed model has multiple tuning strategies to enhance predictive accuracy while preventing overfitting. Neural network hyperparameters with learning rate, dropout, filter sizes, and batch size were optimized experimentally, with adaptive callbacks like ReduceLROnPlateau and EarlyStopping facilitating efficient training. The Random Forest was fine-tuned by using 1000 trees, limiting tree depth to 50, and selecting random feature subsets. Ensemble weights were adjusted to 0.7 for RF and 0.3 for the neural network after testing various combinations. However, these optimizations ensured the hybrid model maintained stability and strong generalization.
Reasoning for choosing the combination shown in Table 3.2:

3.4.9 K-Fold Cross-Validation (Stratified)

K-fold cross-validation splits the dataset into K equal parts (folds), preserving class proportions in each fold.

Table 3.2: Reasoning for choosing combination for Hybrid model

Sequence	CNN+DNN	CNN+DNN+Embedding	Add Attention	Add only PathA (NN)	Add only PathB (RF)	Proposed model
Accuracy	0.80	0.82	0.84	0.83	0.82	0.87
Macro avg						
Precision	0.85	0.88	0.88	0.76	0.80	0.87
Recall	0.69	0.69	0.74	0.72	0.71	0.78
F1	0.68	0.70	0.76	0.72	0.72	0.80
Weighted avg						
Precision	0.81	0.84	0.85	0.82	0.83	0.87
Recall	0.80	0.82	0.84	0.83	0.82	0.87
F1	0.79	0.81	0.84	0.82	0.81	0.86

- In K-fold cross validation, the main data split into K parts.
- The model will be trained into K-1 parts and tasted on the remaining ones.
- This will be repeated K times so that each part is tested and the average of all the results like accuracy,precision gives a more reliable performance estimate.

Benefits:

- It reduces the bias from a single train/test split.
- Provides reliable estimates, especia;;y when the classes are uneven.
- It also ensures fair comparison across all models.

3.4.10 Evaluation Metrics

Accuracy: Accuracy measures how often the models predictions are correct, how many times the model guesses, whether its positive or negative, actually matched the true outcomes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.31}$$

Accuracy is most effective when the dataset is balanced.

Precision: Precision shows how many of the cases the model labeled as positive were actually correct. In short it tells us how reliable the models' positive predictions are.

$$Precision = \frac{TP}{TP + FP} \tag{3.32}$$

useful when making a wrong positive prediction has serious issues.

Recall: Recall shows how good the model is at finding all actual positive cases. In short it measures how many of the real positives the model successfully found.

$$Recall = \frac{TP}{FN + TP} \tag{3.33}$$

important when missing a positive case could lead to serious effects.

F1-Score: The F1-score is the combination of precision and recall into a single value to show how well the model is balanced in accuracy and coverage. The F1-score especially works when one type of result is less common and gives a fair overall measure of the model's performance.

$$F1\text{-}Score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall} \right) \tag{3.34}$$

Confusion Matrix: A confusion matrix is like a table of content which shows where the model made correct predictions and where it made mistakes for each of the classes. It lines up the models predictions against the true labels that's why it's easy to see the correct hits or misses.
First, collect true labels and the models predicted labels from the test data.
Make a table where the rows show the real classes and the columns show the predicted class.
All the diagonal boxes show correct predictions, and others show where the model gets confused.
From the table, we can calculate precision,recall and F1-score to see how well the model's performance and guide fixes.

ROC Curve: The ROC curve is like a graph where it shows how well a model separates two classes when the decision threshold changes. It also shows the balance between true positives(TPR), mistakenly marking negatives as positives(FPR) across all possible thresholds.

Key Terms & Equations:

True Positive Rate (TPR) / Recall:

$$TPR = \frac{TP}{FN + TP} \tag{3.35}$$

False Positive Rate (FPR):

$$FPR = \frac{FP}{TN + FP} \tag{3.36}$$

X-axis: False Positive Rate (FPR)
Y-axis: True Positive Rate (TPR)
Each point on the ROC curve corresponds to a different threshold.

AUC (Area Under the Curve):

$$AUC = \int_0^1 TPR(FPR)\, d(FPR) \tag{3.37}$$

Use the ROC/AUC to compare models without depending on one fixed threshold. Especially when the cost ratio is not extreme and the positive/negative class distribution is stable.
If one class is much smaller or the cost errors are very uneven then it's better to check the precision recall curve for a clear picture.

3.4.11 Explainability Methods

Explainability Methods used to observe the reasoning behind the model's true or false prediction.
In our study, we use the SHAPley Additive exPlanation (SHAP) forced plot to observe the contribution and pattern of pollutant features towards the hybrid model's output as well as the prediction. It uses Shapley values to recognize the pattern.
In the force plot, the base value is the average model output across the training dataset. Each feature positions according to this base value upward or downward toward the predicted value.
Features shown in red (or pushing to the right) have positive SHAP values, which means those features contribute more to predicting class. Features in blue (or pushing to the left) mean they decrease the predicted probability. The length of the bars or arrows represents the strength of each feature's impact. The longer the bar, the stronger its influence.
This visualization is a very effective way to understand how the model actually predicts true or false classes.

Chapter 4

Result and Discussion

4.1 Accuracy:

Fig. 4.1 shows Hybrid model accuracy is 0.87. Random forest, LightGBM, and CatBoost are in the 2nd highest position with 0.84 accuracy. LSTM 0.71, SVM 0.62, and GRU 0.59.

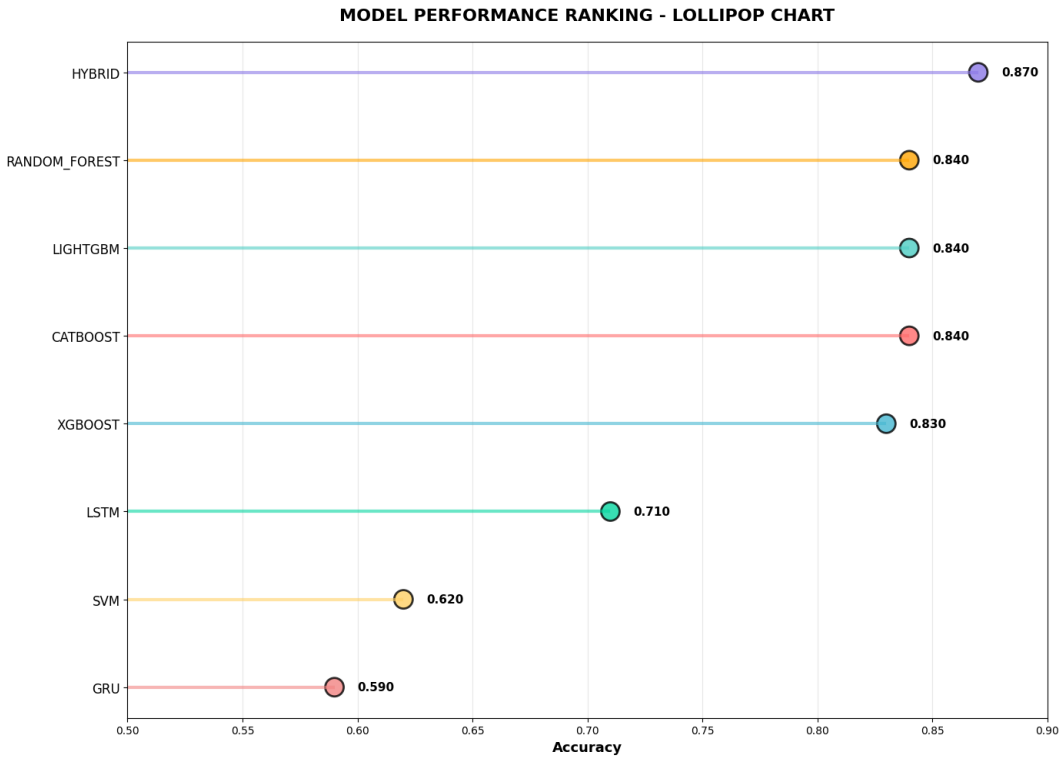


Figure 4.1: Model Accuracy comparison.

4.2 Precision, Recall, F1 score result and analysis:

Machine learning model comparison in Table 4.1, Deep learning model comparison in Table 4.2, and 5-fold K-cross validation in Table 4.3

Table 4.1: Class-wise Precision, Recall, and F1-score for Different Models

Model	**Class**	**Precision**	**Recall**	**F1-score**
CatBoost	Caution	0.80	0.90	0.85
	Good	0.85	0.74	0.79
	Moderate	0.94	0.87	0.90
	Unhealthy	0.81	0.79	0.80
	Very Unhealthy	1.00	0.64	0.78
LightGBM	Caution	0.82	0.89	0.86
	Good	0.80	0.85	0.82
	Moderate	0.94	0.87	0.90
	Unhealthy	0.83	0.77	0.80
	Very Unhealthy	0.83	0.71	0.77
Random Forest	Caution	0.82	0.88	0.85
	Good	0.82	0.74	0.78
	Moderate	0.91	0.88	0.90
	Unhealthy	0.82	0.79	0.81
	Very Unhealthy	0.77	0.71	0.74
SVM	Caution	0.56	0.83	0.67
	Good	0.74	0.53	0.62
	Moderate	0.77	0.25	0.38
	Unhealthy	0.68	0.70	0.69
	Very Unhealthy	0.67	0.29	0.40
XGBoost	Caution	0.81	0.87	0.84
	Good	0.81	0.76	0.78
	Moderate	0.87	0.85	0.86
	Unhealthy	0.82	0.81	0.82
	Very Unhealthy	0.87	0.50	0.63

Table 4.2: Class-wise Precision, Recall, and F1-score for GRU, LSTM, and Hybrid Models

Model	Class	Precision	Recall	F1-score
GRU	**Accuracy**		**0.59**	
	Caution	0.57	0.73	0.64
	Good	0.65	0.53	0.58
	Moderate	0.79	0.34	0.47
	Unhealthy	0.59	0.70	0.64
	Very Unhealthy	0.33	0.21	0.26
LSTM	**Accuracy**		**0.71**	
	Caution	0.75	0.74	0.75
	Good	0.67	0.53	0.59
	Moderate	0.77	0.69	0.73
	Unhealthy	0.65	0.83	0.73
	Very Unhealthy	0.56	0.36	0.43
Hybrid	**Accuracy**		**0.87**	
	Caution	0.83	0.91	0.87
	Good	0.85	0.89	0.87
	Moderate	0.95	0.90	0.92
	Unhealthy	0.87	0.84	0.86
	Very Unhealthy	0.83	0.36	0.50

5-fold K-cross validation:

Table 4.3: Overall performance comparison of models using 5-fold cross-validation.

Model	Accuracy	Precision	Recall	F1-score
CatBoost	0.8271 ± 0.0125	0.8398 ± 0.0431	0.7504 ± 0.0336	0.7776 ± 0.0373
LightGBM	0.8428 ± 0.0096	0.8426 ± 0.0170	0.8262 ± 0.0256	0.8331 ± 0.0212
SVM	0.6144 ± 0.0293	0.6657 ± 0.0364	0.6144 ± 0.0293	0.5919 ± 0.0474
XGBoost	0.8388 ± 0.0071	0.8278 ± 0.0156	0.7723 ± 0.0268	0.7936 ± 0.0231
Random Forest	0.8485 ± 0.0036	0.8524 ± 0.0166	0.7824 ± 0.0219	0.8085 ± 0.0177
GRU	0.6095 ± 0.0147	0.6499 ± 0.0337	0.5493 ± 0.0205	0.5753 ± 0.0241
LSTM	0.6229 ± 0.0300	0.6555 ± 0.0478	0.5601 ± 0.0243	0.5855 ± 0.0335
Hybrid	0.8520 ± 0.0251	0.8446 ± 0.0262	0.8430 ± 0.0251	0.8410 ± 0.0256

4.3 ROC-AUC result and comparison:

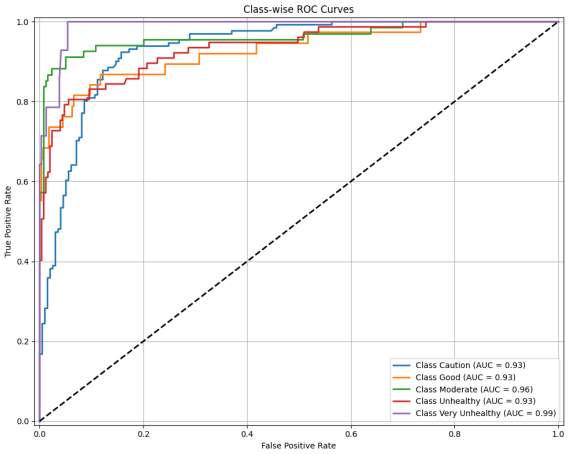


Figure 4.2: ROC curve (Catboost).

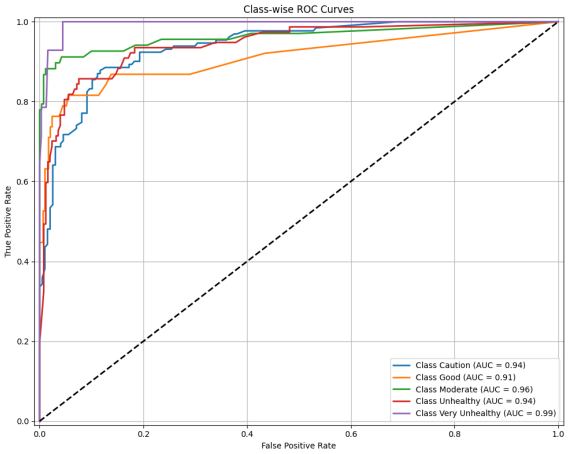


Figure 4.3: ROC curve (Random forest).

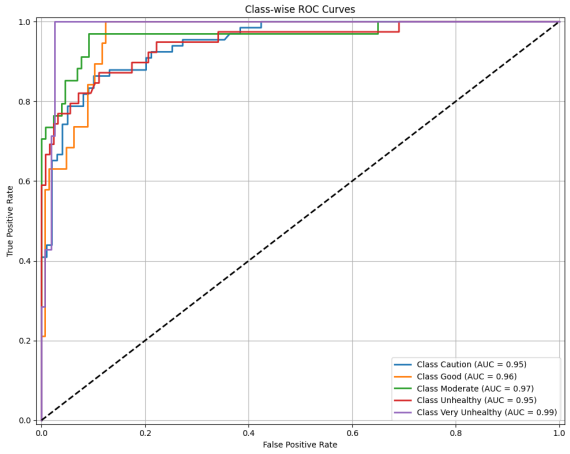


Figure 4.4: ROC curve (Lightgbm).

For CatBoost (Figure 4.2), the ROC curve shows quite good performance across all classes. The AUC scores are 0.93 for Caution, 0.93 for Good, 0.96 for Moderate,

0.93 for Unhealthy, and 0.99 for Very Unhealthy. All of them indicate a very good ROC curve.
For Random Forest (Figure 4.3), the ROC curve is also very strong. The AUC scores are 0.94 for Caution, 0.91 for Good, 0.96 for Moderate, 0.94 for Unhealthy, and 0.99 for Very Unhealthy. This confirms a very good ROC curve across all classes.
For LightGBM (Figure 4.4), the ROC curve shows a very good class separation. The AUC values are 0.95 for Caution, 0.96 for Good, 0.97 for Moderate, 0.95 for Unhealthy, and 0.99 for Very Unhealthy. This confirms a strong ROC performance.

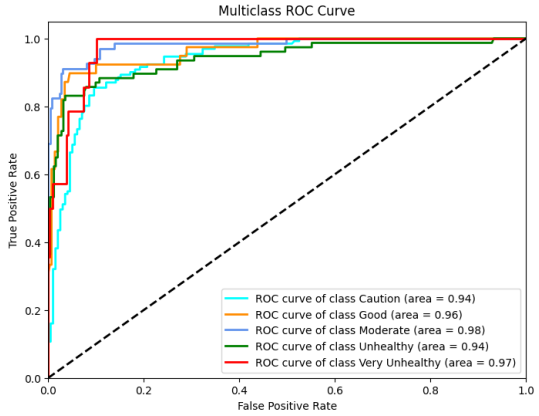


Figure 4.5: ROC curve (XGboost).

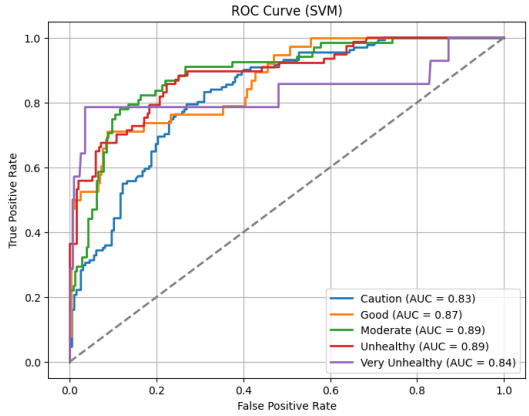


Figure 4.6: ROC curve (SVM).

For XGBoost (Figure 4.5), the ROC curve accuracy is really strong. The AUC values are 0.94 for Caution and Unhealthy, while Good is 0.96, Moderate is 0.98, and Very Unhealthy is 0.97. This reflects excellent ROC performance.
For SVM (Figure 4.6), the ROC curve shows moderate to good discrimination. The AUC scores are 0.83 for Caution, 0.87 for Good, 0.89 for Moderate, 0.89 for Unhealthy, and 0.84 for Very Unhealthy, which shows that the ROC curve is good but not excellent.

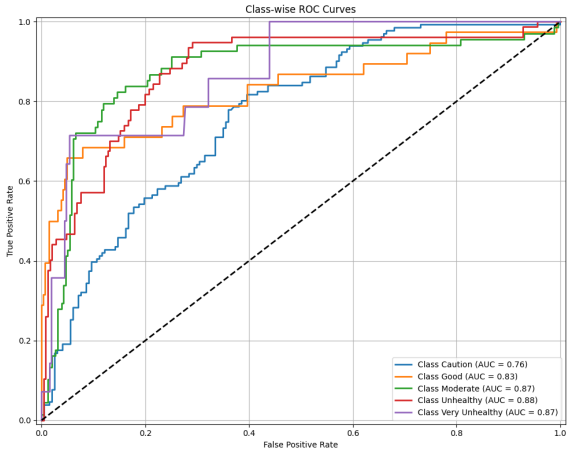


Figure 4.7: ROC curve (GRU).

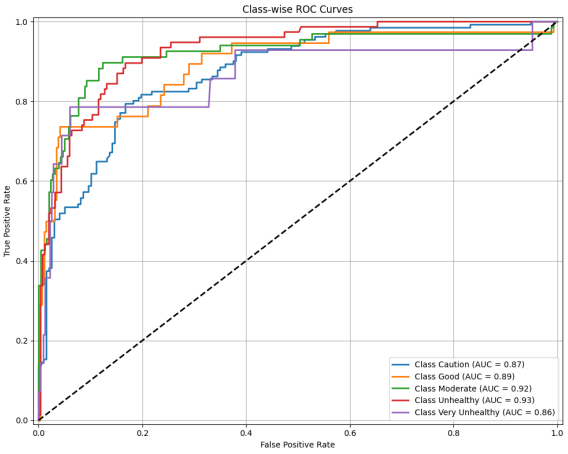


Figure 4.8: ROC curve (LSTM).

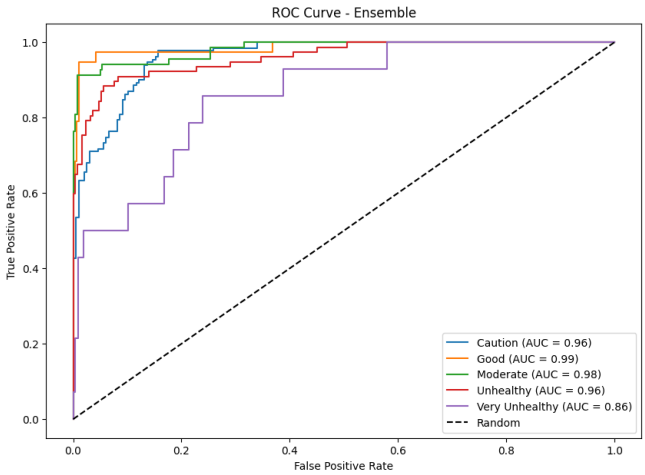


Figure 4.9: ROC curve (Hybrid).

For GRU Fig. 4.7, the ROC curve demonstrates fair to good performance. The AUC scores are 0.76 for Caution, 0.83 for Good, 0.87 for Moderate, 0.88 for Unhealthy, and 0.87 for Very Unhealthy. It shows that the ROC curve is acceptable but weaker for Caution.
For LSTM Fig. 4.8, the ROC curve shows good performance overall. The AUC values are 0.87 for Caution, 0.89 for Good, 0.92 for Moderate, 0.93 for Unhealthy, and 0.86 for Very Unhealthy. This demonstrates that the ROC is good, with slightly lower performance on Very Unhealthy.
For the Hybrid model Fig. 4.9, the ROC curve is highly accurate. The AUC scores are 0.96 for Caution, 0.99 for Good, 0.98 for Moderate, 0.96 for Unhealthy, and 0.86 for Very Unhealthy, showing a very good ROC curve, though slightly weaker for Very Unhealthy.

4.4 Comparison Table

Table 4.4: AUC values for different models and AQI classes.

Model	Good	Caution	Moderate	Unhealthy	Very Unhealthy
Machine Learning					
CatBoost	0.93	0.93	0.96	0.93	0.99
Random Forest	0.91	0.94	0.96	0.94	0.99
LightGBM	0.96	0.95	0.97	0.95	0.99
XGBoost	0.96	0.94	0.98	0.94	0.97
SVM	0.87	0.83	0.89	0.89	0.84
Deep Learning					
GRU	0.83	0.76	0.87	0.88	0.87
LSTM	0.89	0.87	0.92	0.93	0.86
Hybrid	0.99	0.96	0.98	0.96	0.86

4.5 Confusion Matrix result and analysis:

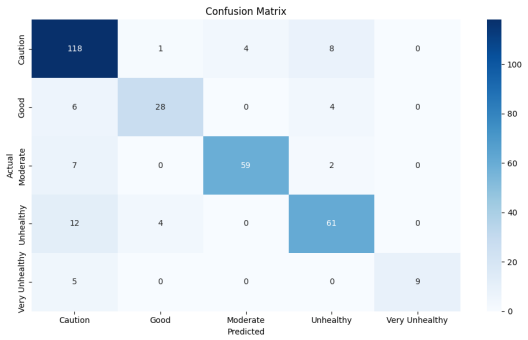


Figure 4.10: Catboost confusion matrix.

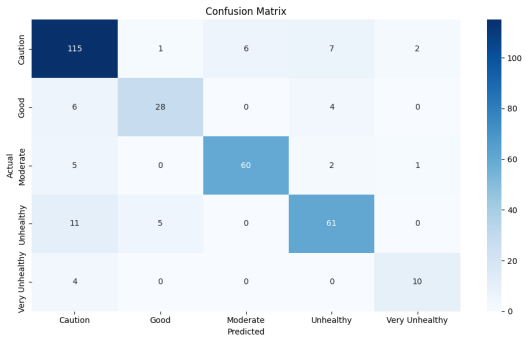


Figure 4.11: Random Forest confusion matrix.

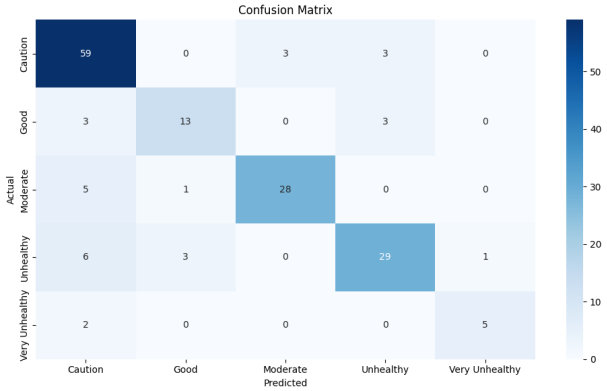


Figure 4.12: LightGBM confusion matrix.

In the figure, each row of the matrix represents the true class, while each column corresponds to the predicted class. Correct predictions are displayed along the diagonal.

In Fig. 4.10, we can see the CatBoost model shows a very strong predictive accuracy across the five air quality categories, especially performing well in identifying 'Caution', 'Unhealthy', and 'Good'. But , it sometimes misclassifies more severe categories as milder categories, such as unclear 'Very Unhealthy' with 'Caution'. Because some classes have fewer samples the model slightly underestimates severity.

In Fig. 4.11, the Random Forest model displays steady performance, properly predicting 'Caution', 'Moderate', and 'Unhealthy' categories. Although the model is reliable, it often underestimates the air quality levels. The 'unhealthy' and 'very unhealthy' cases are predicted as 'caution'. This shows that the model could not do well with less common classes.

In Fig. 4.12, the LightGBM model performs fairly well, accurately predicting 'Caution', 'Moderate', and 'Unhealthy' categories but showing lower accuracy in 'Good' and 'Very Unhealthy'. Mostly blunders between nearby severity levels, shows that the model understands main patterns but has some trouble in separating the categories clearly.

Overall, by analyzing the CatBoost model, it gives the best performance, showing strong accuracy and better difference among air-quality categories. The Random Forest model comes very close, providing stable but somewhat cautious predictions that often underestimate high-severity cases. LightGBM works very well, but its performance is a bit lower and it gets into trouble while there are similar air quality levels. All the models can improve their performance by balancing their classes, adjusting the model settings, and improving features, which helps to separate nearby categories more clearly.

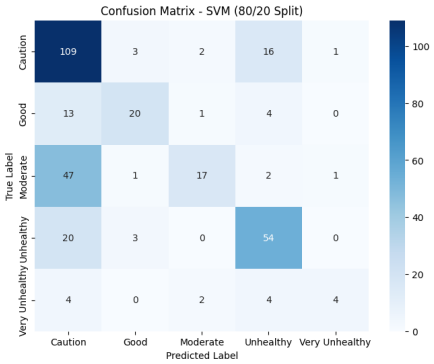


Figure 4.13: SVM confusion matrix.

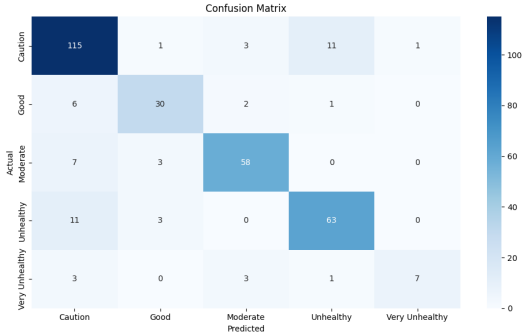


Figure 4.14: XGboost confusion matrix.

In Fig. 4.13, the SVM classifier is well represented in enumerating the 'Caution' and 'Unhealthy' categories but has inconvenience with adjoining and less narrated classes. It often miscategorized average and high-severity categories as safer ones, representing a conservative aptitude. The lower recall for scarce categories like 'Very Unhealthy' is evidence that the model is impressionable to class asymmetry and overlapping characteristic ambients.

In Fig. 4.14, the XGBoost model unravels rigid and well-balanced alignment perfection, exactly placing the 'Caution,', 'Moderate', and 'Unhealthy' categories. Most misclassifications happen between contiguous classes, and while it conducts the ordered nature of brutality better than SVM, it still contends with infrequent categories such as 'Very Unhealthy'. Its nourished design helps it assume complicated patterns more successfully.

In general, the XGBoost model omits SVM in classifying air-quality categories, paying off supreme precision and an improved balance between exactness and retraction. The SVM model neglects to commit to a secured, less harsh prognosis, which marks its efficiency in finding troublesome ambiences such as 'Very Unhealthy'. In contrast, XGBoost does a better job separating between the categories, though both models could still be developed by rebalancing the data and enhancing feature separability to better distinguish brutality levels.

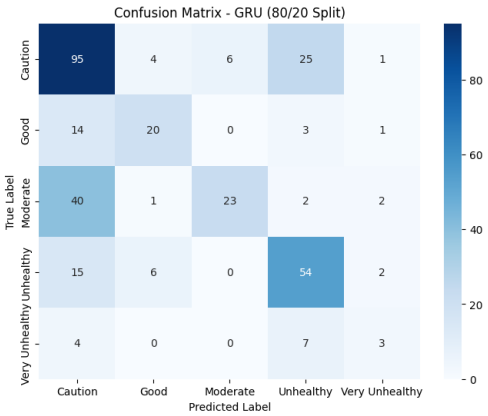


Figure 4.15: GRU confusion matrix.

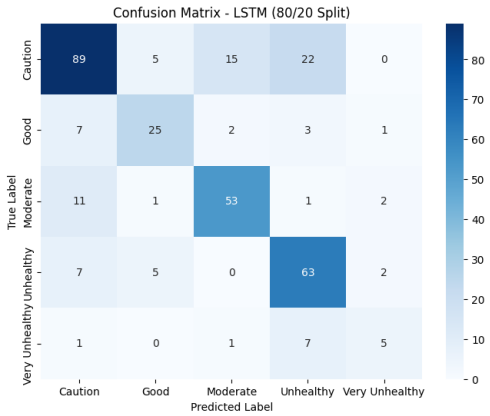


Figure 4.16: LSTM confusion matrix.

In Fig. 4.15, we can see the GRU model performs well in predicting the 'Caution' and 'Unhealthy' categories but has faced some difficulties with 'Moderate' and 'Very Unhealthy'. Offently it guesses the air quality as less serious and predicts severe or moderate cases as safer ones. So we can assume that it is affected by uneven class size and same features between nearby categories. It's useful for the common classes but faces difficulties in rare and more severe categories.
In Fig. 4.16 LSTM model shows powerful overall performance, accurately identifying 'Caution,' 'Moderate,' and 'Unhealthy' categories. Most mistakes occur in nearby classes. Shows that the model can learn time patterns even when some categories overlap. It can still predict some severe cases as less serious, but always balanced accuracy across most classes and shows that it can learn time based patterns better than the GRU model.
Overall, the LSTM model works better than the GRU in classifying air quality levels. Both models can detect common classes but LSTM separates the levels more accurately and gives a very balanced result. While GRU is fast and sometimes underestimates severe cases and is more affected by the uneven class size. The models can be improved by balancing the data and some features to get better detection in rare and extreme air quality conditions.

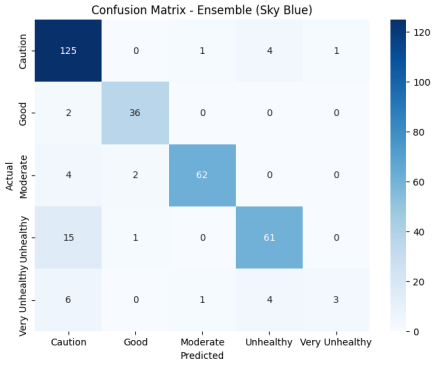


Figure 4.17: Hybrid confusion matrix.

In Fig. 4.17, the hybrid model shows the most powerful overall performance, achieving high accuracy for the 'Caution,' 'Moderate,' and 'Unhealthy' categories, with most predictions accurately aligned along the diagonal. The good results show that

the model combines features well and makes better decisions than the individual models. But it is unlikely it has trouble with the 'very unhealthy' category because there are some samples which are in extreme conditions and most of the mistakes happen between nearby levels which seems neutral rather than random. Overall, the model gives balanced and stable predictions, doing better than single models.
Among the eight models, the Hybride model performs the best. The model combines the capabilities of deep learning, which is responsible for recognizing time patterns and strong decision making of ensemble methods. The LSTM shows good balance in accuracy and has the ability to learn from the time-based data.
CatBoost, XGBoost, and Random Forest also perform well, in common air quality levels, but struggle in the rare and severe cases. Where SVM and GRU often guess the air quality is less serious but give safer predictions. LightGBM is very fast but not strong in classifications and gets stuck on similar categories. At last the hybrid model is the most reliable and adaptable. Making fewer mistakes between nearby classes and getting high accuracy even if there are some classes that have fewer samples.

4.6 Explainability with SHAP

4.6.1 Force plot for correctly predicted classes

Unhealthy Class:

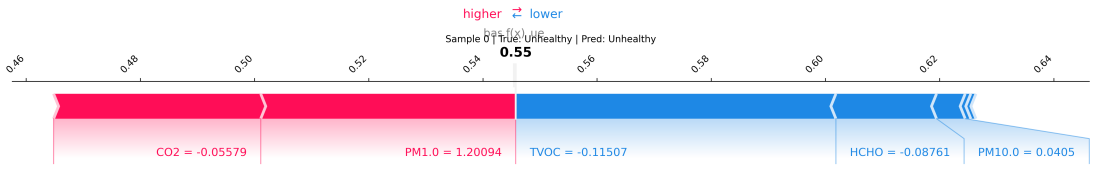


Figure 4.18: True class unhealthy predicted class unhealthy

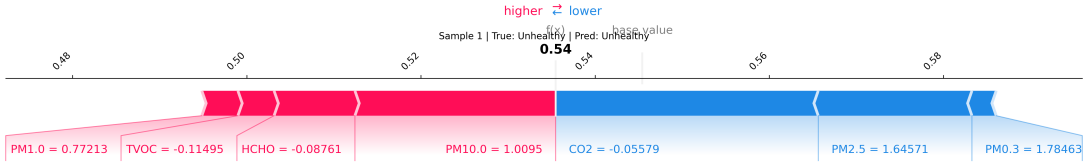


Figure 4.19: True class unhealthy predicted class unhealthy

For Fig. 4.18 and Figure 4.19, base values are 0.55 and 0.54, respectively, which correctly predict the class as unhealthy. In sample Figure 4.18, CO_2 and $PM_{1.0}$ are in higher positions, which push the model upwards to the predicted class. Among them, $PM_{1.0}$ is more responsible than CO_2. On the other hand, for Figure 4.19, PM_1, TVOC, HCHO, and PM_{10} are in higher positions, which push the model towards the predicted class. Here PM_{10} is in the highest position.

Alternatively, in Fig. 4.18 TVOC, HCHO, and PM_{10} are in lower positions, which pull the model downwards to the predicted class; among them TVOC plays a major role as a decreaser. Moreover, in Figure 4.19, CO_2, $PM_{2.5}$, and $PM_{0.3}$ pull the model downwards to the predicted class; among them, CO_2 plays a greater role.

Caution Class:

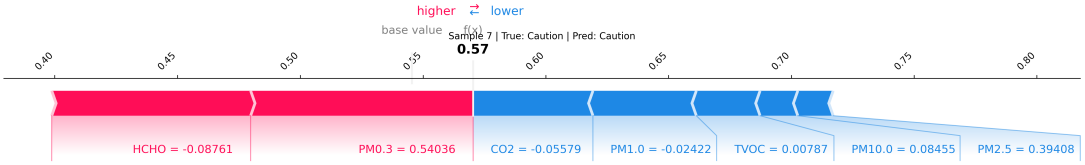


Figure 4.20: True class caution predicted class caution

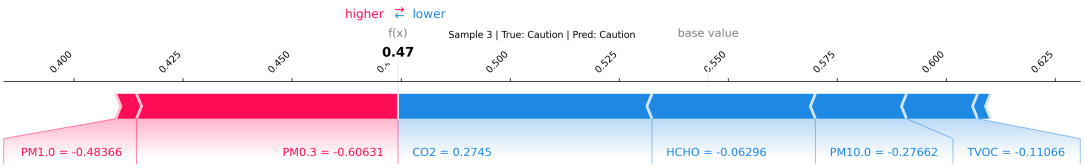


Figure 4.21: True class caution predicted class caution

For Fig. 4.20 and Figure 4.21, base values are 0.57 and 0.47, respectively, which correctly predict the class as Caution. In Figure 4.20, HCHO and $PM_{0.3}$ are in higher positions, which push the model upwards to the predicted class. Among them, $PM_{0.3}$ shows a greater length of bar. On the other hand, for Figure 4.21, $PM_{1.0}$ and $PM_{0.3}$ are in higher positions, which push the model towards the predicted class. Here $PM_{0.3}$ shows a greater length.
Alternatively, in Figure 4.20, CO_2, $PM_{1.0}$, TVOC, PM_{10}, and $PM_{2.5}$ are in lower positions, which pull the model downwards to the predicted class; among them, CO_2 and $PM_{1.0}$ show the largest bar length. Moreover, in Figure 4.21, CO_2, HCHO, PM_{10}, and TVOC pull the model downwards to the predicted class; among them, CO_2 and HCHO play a greater role.

Good Class:

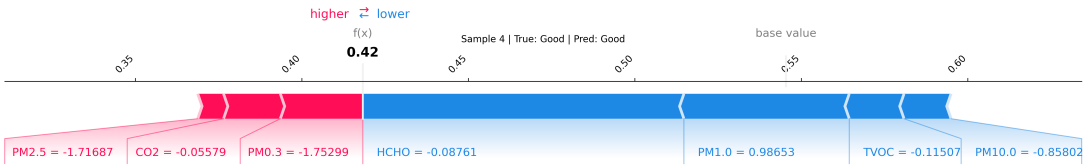


Figure 4.22: True class good predicted class good

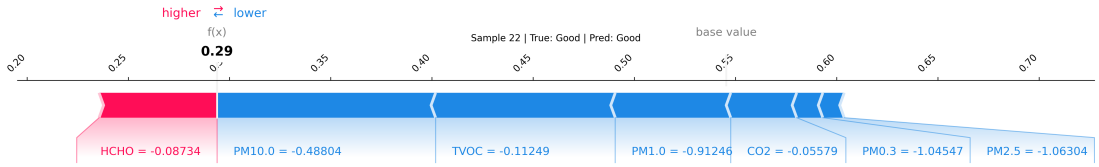


Figure 4.23: True class good predicted class good

For Fig. 4.22 and Figure 4.23, base values are 0.42 and 0.29, respectively, which correctly predict the class as good. In Figure 4.22, CO_2, $PM_{2.5}$, and $PM_{0.3}$ are in higher positions, which push the model upwards to the predicted class. Here, $PM_{0.3}$ is more responsible. On the other hand, for Figure 4.23, HCHO is in a higher position, which pushes the model towards the predicted class.
Alternatively, in Fig. 4.22 TVOC, HCHO, $PM_{1.0}$, and PM_{10} are in lower positions, which pull the model downwards to the predicted class; among them HCHO plays a major role as a decreaser. Moreover, in Figure 4.23, CO_2, $PM_{2.5}$, PM_{10}, TVOC, $PM_{1.0}$, $PM_{2.5}$, and $PM_{0.3}$ pull the model downwards to the predicted class; among them, PM_{10} plays a greater role.

Moderate Class:

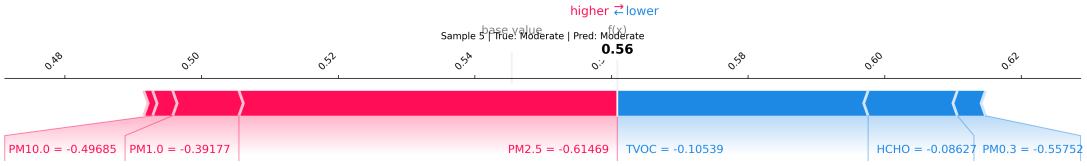


Figure 4.24: True class moderate predicted class moderate

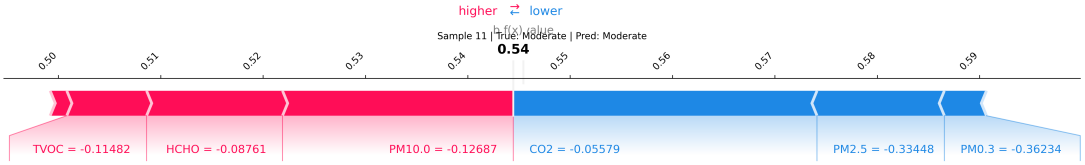


Figure 4.25: True class moderate predicted class moderate

For Fig. 4.24 and Figure 4.25, base values are 0.56 and 0.54, respectively, which correctly predict the class as moderate. In Figure 4.24, PM_{10}, $PM_{2.5}$, and $PM_{1.0}$ are in higher positions, which push the model upwards to the predicted class. Among them, $PM_{2.5}$ is more responsible than CO_2. On the other hand, for Figure 4.25, PM_{10}, TVOC, and HCHO are in higher positions, which push the model towards the predicted class. Here PM_{10} is in the highest position.

Alternatively, in Fig. 4.24, $PM_{1.0}$, $PM_{2.5}$, and PM_{10} are in lower positions, which pull the model downwards to the predicted class; among them, $PM_{2.5}$ plays a major role as a decrease. Moreover, in Figure 4.25, CO_2, $PM_{2.5}$, and $PM_{0.3}$ pull the model downwards to the predicted class; among them, CO_2 plays a greater role.

Very unhealthy class:

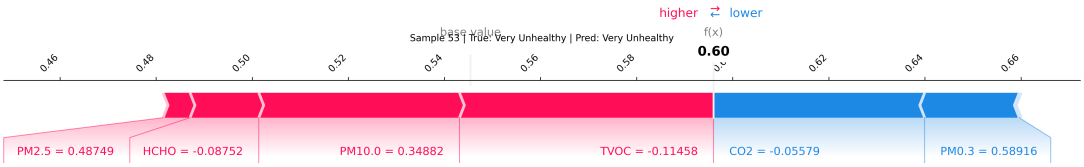


Figure 4.26: True class very unhealthy predicted class very unhealthy

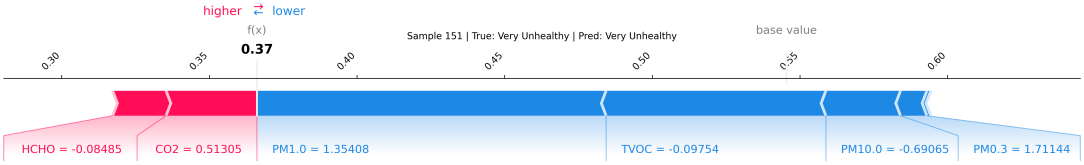


Figure 4.27: True class very unhealthy predicted class very unhealthy

For Fig. 4.26 and Fig. 4.27, base values are 0.60 and 0.37, respectively, which correctly predict the class as very unhealthy. In Fig. 4.26, HCHO, $PM_{2.5}$, PM_{10}, and TVOC are in higher positions, which push the model upwards to the predicted class. Among them, TVOC shows a greater length of bar. On the other hand, for Figure 4.27, HCHO and CO_2 are in higher positions, which push the model towards the predicted class. Here CO_2 shows a greater length.
Alternatively, in Fig. 4.26, CO_2 and $PM_{0.3}$ are in lower positions, which pull the model downwards to the predicted class; among them, CO_2 shows the largest bar length. Moreover, in Fig. 4.27, $PM_{1.0}$, $PM_{0.3}$, PM_{10}, and TVOC pull the model downwards to the predicted class; among them, $PM_{1.0}$ plays a greater role.//

4.6.2 Force plot for error prediction

Very Unhealthy:

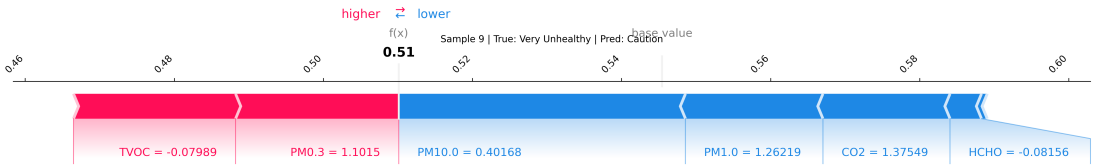


Figure 4.28: True class very unhealthy predicted class caution

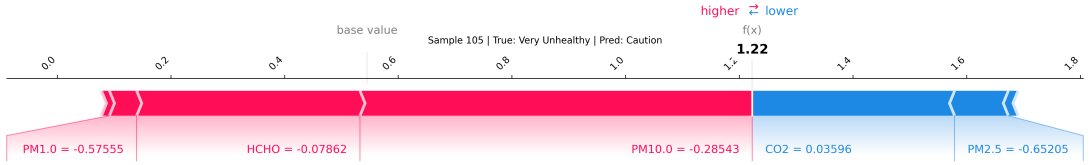


Figure 4.29: True class very unhealthy predicted class caution

For Fig. 4.28 and Fig. 4.29, model outputs are 0.51 and 1.22, respectively, and predict the class caution, but the true class is very unhealthy. In Fig. 4.28, TVOC and $PM_{0.3}$ are in higher positions, which push the model upwards to the predicted class. On the other hand, for Fig. 4.29, $PM_{1.0}$, PM_{10}, and HCHO are in higher positions, which push the model towards the predicted class. Here PM_{10} shows a greater magnitude.
Alternatively, in Fig. 4.28, CO_2, $PM_{1.0}$, HCHO, and PM_{10} are in lower positions, which pull the model downwards to the predicted class; among them, PM_{10} shows the largest magnitude. Moreover, in Fig. 4.29, CO_2 and $PM_{2.5}$ pull the model downwards to the predicted class; among them, CO_2 shows a greater magnitude.

Unhealthy class:

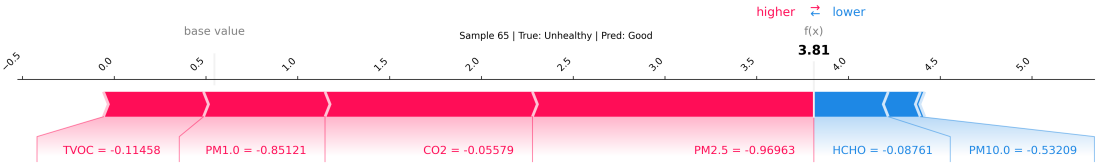


Figure 4.30: True class unhealthy predicted class good

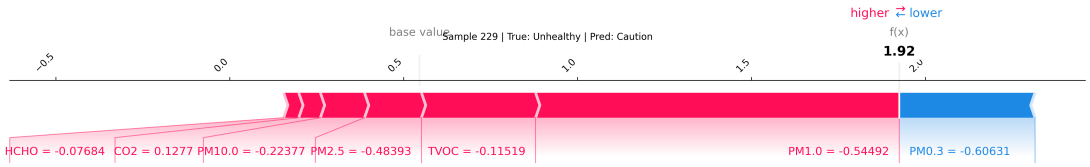


Figure 4.31: True class unhealthy predicted class caution

For Figure 4.30 and Figure 4.31, model outputs are 3.81 and 1.92, with predicted classes 'good' and 'caution', respectively, but the true class is unhealthy. In Figure 4.30, TVOC, $PM_{1.0}$, CO_2, and $PM_{2.5}$ are in higher positions, which push the model upwards to the predicted class. On the other hand, for Figure 4.31, HCHO, CO_2, $PM_{2.5}$, PM_{10}, TVOC, and are in higher positions, which push the model towards the predicted class. Here $PM_{1.0}$ shows a greater magnitude.

Alternatively, in sample Figure 4.30, HCHO and PM_{10} are in lower positions, which pull the model downwards to the predicted class; among them, HCHO shows the largest magnitude. Moreover, in Figure 4.31, $PM_{0.3}$ pulls the model downwards to the predicted class; among them, HCHO shows a greater magnitude.

Caution class:

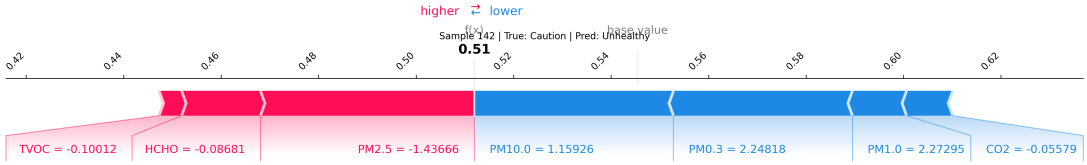


Figure 4.32: True class caution predicted class unhealthy

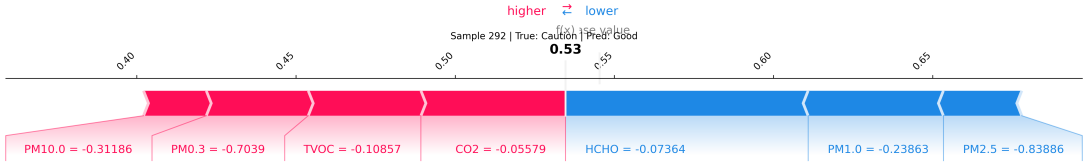


Figure 4.33: True class caution predicted class good

For Figure 4.32 and Figure 4.33, model outputs are 0.51 and 0.53, with predicted classes unhealthy and good, respectively, but the true class is caution. In Figure 4.32, TVOC, HCHO and $PM_{2.5}$ are in higher positions, which push the model upwards to the predicted class. On the other hand, for Figure 4.33, $PM_{0.3}$, PM_{10}, TVOC, and CO_2 are in higher positions, which push the model towards the predicted class. Here CO_2 shows a greater magnitude.
Alternatively, in sample Figure 4.32, CO_2, $PM_{0.3}$, $PM_{0.3}$, $PM_{1.0}$, and PM_{10} are in lower positions, which pull the model downwards to the predicted class; among them, PM_{10} shows the largest magnitude. Moreover, in Figure 4.33, HCHO, PM_{10}, and $PM_{2.5}$ pull the model downwards to the predicted class; among them, HCHO shows a greater magnitude.

Moderate class:

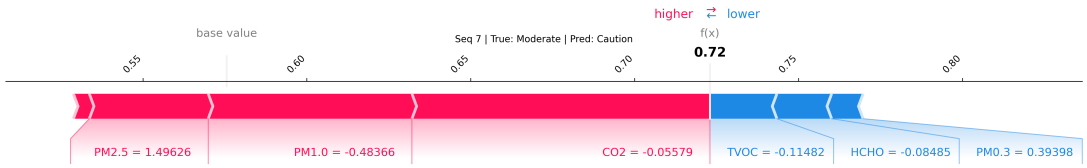


Figure 4.34: True class moderate predicted class caution

For Figure 4.34, the model output is 0.72 with prediction caution, but the true class is moderate. PM_{10}, CO_2, and $PM_{2.5}$ are in higher positions, which push the model upwards to the predicted class. CO_2 shows the highest magnitude.
Alternatively, TVOC, HCHO, and $PM_{0.3}$ are in lower positions, which pull the model downwards to the predicted class; among them, TVOC shows the largest magnitude.

Good class:

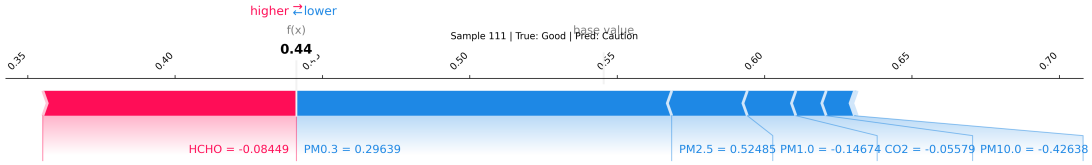


Figure 4.35: True class good predicted class caution

For Figure 4.35, the model output is 0.44 with prediction caution, but the true class is good. HCHO is in a higher position, which pushes the model upwards to the predicted class.
Alternatively, $PM_{2.5}$, $PM_{1.0}$, PM_{10}, CO_2, and $PM_{0.3}$ are in lower positions, which pull the model downwards to the predicted class; among them, $PM_{0.3}$ shows the largest magnitude.
Analysis: The error prediction occurs for feature overlap, marginal pollutant differences, and competing positive–negative contributions. However, the class imbalance is the main issue for the model error prediction.

Chapter 5

Discussion

5.1 Model Behavior Interpretation

The reason for the better performance in ML models, Gradient boosting models such as CatBoost,LightGBM and XGBoost performed better because they learn from their previous tree mistakes, which helps them to become more accurate,faster and efficient and also can handle complex pollutant patterns effectively. The LightGBM model grows trees leaf by leaf. On the other hand CatBoost uses balanced trees, both models allowing to capture small but effective relationships between pollutant levels and AQI categories. The regularization and the early stopping part helped to prevent overfitting, which is very much important because sometimes air quality data can have sudden changes or noisy readings. Besides, the random forest model worked well by using randomization to create diverse trees, which guides to reduce error and keeps the result stable without overfitting.
The key features like $PM_{2.5}$, PM_{10}, CO and other gases played a vital role in predicting the AQI. All the tree based models effectively split nodes using the most informative features to separate AQI classes. SVM performed well when the data was balanced and the feature was scaling properly, it could not capture the non linear relationship as the boosting models did.
The reason for the lower performance in ML models, SVM had faced trouble while handling complex relationships between pollutant features, even when it was using the RBF kernel. The model also has some issues regarding scaling and outliers, meaning very high or uneven values of pollutants make it harder to find the best separating line between class to class.
For the deep learning models, LSTM(0.71) performed better than GRU(0.59) and our hybrid model combining features performed the best(0.87). The main reason for LSTM better performance is better than GRU because its memory cells and three gates helped it to remember both long term and short term patterns in the dataset. On the other hand, the Hybride model is the combination of deep learning and ML, the deep learning part is used for feature embeddings and the ML part is used for refined classification. This allows it to capture complex patterns and improve its accuracy. Both LSTM and hybrid models can capture time based variations in pollutants, such as sudden spikes or gradual changes in $PM_{2.5}$ and CO levels. GRU has a simple structure and struggles with longer dependencies.

5.2 Implications for Students

By knowing the Air Quality Index (AQI) and the pollutant levels, all the students can take some necessary steps to protect their health, such as wearing masks and using air purifiers when the air quality is poor. Also, they can plan their daily activities wisely by avoiding outdoor exercise and choosing cleaner times for walking. For students this system helps them to understand real-world pollution patterns and can apply predictive models. Overall, it increases health awareness, makes them more efficient in their daily planning and inculcates environmentally responsible habits among students. In short, it can be said that AQI forecasts can be used by students to stay healthy and practical insights can be gained by them from data-driven models.

5.3 System Flexibility

The AQI prediction system is flexible and can be used in different campuses or in other urban areas with a minimal change. Collecting air quality data from locals, the models can be restrained to match the specific pollution patterns of each location. Our system also allows for adding new sensors to cover a larger area and can handle different types of pollutants in different environments. The daily planning suggestions can be customized based on local preference. Models like LightGBM, CatBoost, LSTM can quickly adapt to new data, ensure accurate predictions and very useful recommendations wherever it is deployed.

Chapter 6

Limitations

Even though the study had encouraging findings, different limitations confine its scope and ought to be taken into consideration. These involve the dataset, model choice, system design, and ethics.

6.1 Data Limitations

The dataset was constrained in terms of geographic area and pollutant coverage. Though the MT15 sensor measured primary indicators such as CO_2, CO, TVOC, HCHO, and particulate matter ($PM_{0.3}$, $PM_{1.0}$, $PM_{2.5}$, PM_{10}), it did not measure other vital pollutants such as O_3, NO_2, or SO_2, required for comprehensive air quality forecasting. Also, data were restricted to a single institutional campus and thus had regional and context-specific bias. Consequently, the findings would not be extendable to rural areas or to urban areas with other pollution sources.
Issues related to devices were also problematic. Although the MT15-Tuya was calibrated to reference-grade monitors, consumer-grade sensor architecture had the potential to introduce noise and variability in measurements. The disproportionate allocation of samples to monitoring locations was a basic limitation as well. For example, certain areas such as basements and traffic paths yielded more results than other less-traveled places such as rooftops or washrooms. This caused uneven data representation in spatial categories and hence unbalanced inputs for classification. Finally, the dataset was small in size (1,637 samples), which restricted more data-hungry models such as GRU and LSTM from delivering their optimal performance because these structures are known to require larger datasets to generalize effectively.

6.2 Model Limitation

- Class imbalance
- Limitation of tuning
- Data overfitting risk

The suggested hybrid deep-learning and ensemble method boosts predictive power at the expense of generalization, interpretability, and simplicity. Future research should

concentrate on utilizing explainability tools (like SHAP), incorporating tabular-specific architectures (like transformer-based or attention-based tabular models), and implementing joint optimization or adaptive ensemble weighting techniques.

6.3 Explainability Constraints

Model explainability was examined with SHAP (Random Forest) and LIME (SVM). Even though these models recognized influential features such as $PM_{2.5}$ and CO_2, they were not without their restrictions. Their computational complexity limited experiments on big data interpretability, and their outputs frequently disagreed. For example, SHAP and LIME occasionally provided inconsistent explanations for the same sample, creating interpretability stability issues. Additionally, the technical nature of these explanations makes them difficult for non-technical stakeholders (e.g., students, policymakers) to readily understand.

6.4 System Constraints

The models were tested and demonstrated in a research environment through Jupyter Notebooks. No attempt was made to implement them as part of an actual real-time application, e.g., a web interface or mobile app. The research is thus overwhelmingly proof-of-concept and thus is not yet capable of facilitating continuous public monitoring or decision aids.

6.5 UX Limitations

Although graphical outputs such as plots, confusion matrices, and rankings of feature importance were produced, there was no documented user experience (UX) testing. The end-user needs (students, instructors, or environmental managers) were not stated explicitly, i.e., usability and understandability of output are untested. Furthermore, the output and documentation were English only, which closed off access to non-English users.

6.6 Ethical and Generalization Concerns

The study deliberately avoided issuing medical or health recommendations, as that is outside its mandate and ethical boundaries. Forecasting was restricted to AQI classes only. AQI criteria are also different geographically—Bangladesh DOE grades differ from US EPA or WHO standards—which raises issues regarding the outcome's generality in other regulation domains. Finally, although ethical clearance was obtained for monitoring on campus, conducting such studies on larger public areas would require more stringent ethical and privacy regulations to protect people and communities.

Chapter 7

Conclusion and Future Work

Air is one of the most essential elements of life, but its quality is dropping day by day because of human activities and natural calamities. Therefore, air quality monitoring has become a global issue. In this paper, we have trained an Air Quality Index (AQI) classifier with different machine learning algorithms and modified it with explainability techniques such as SHAP such that the predictions are explainable and trustworthy. The results showed that machine learning models like CatBoost, LightGBM, Random Forest, and their mix outperformed neural network models remarkably. Of these, the hybrid model achieved the highest overall accuracy, precision, and F1-score.

Our explainability analysis revealed that $PM_{2.5}$, $PM_{0.3}$, CO_2, HCHO, and TVOC were the most significant pollutants for predicting AQI categories. This reflects the influence of both particulate matter as well as gas-type pollutants on air quality. This finding not only confirms that our model performs effectively but also provides environmental agencies with helpful insights to focus on the most dangerous pollutants.

Apart from our success, there were a few limitations in the study, including biased representation of high AQI values, limited data to particular areas and time periods, and computational costs of calculating rich explanations. A growing datasets from different geographic areas and time periods, combining real-time sensor data for continuous monitoring, creating user-friendly dashboards, and mobile apps offering accessible forecasts and health warnings will be required in the future to overcome these limitations.

In conclusion, the combination of explainable AI and machine learning is a strong approach to building such intelligent and reliable air quality monitoring systems. These measures have the potential to provide the opportunity to take necessary measures to protect the environment and human health against the harmful effects of air pollution.

Bibliography

[1] A. K. Majumder, M. N. A. Patoary, A. Al Nayeem, and M. Rahman, "Air quality index (aqi) changes and spatial variation in bangladesh from 2014 to 2019," *Journal of Air Pollution and Health*, 2023.

[2] T. Holloway et al., "Satellite monitoring for air quality and health," *Annual review of biomedical data science*, vol. 4, no. 1, pp. 417–447, 2021.

[3] H. Tariq, F. Touati, D. Crescini, and A. Ben Mnaouer, "State-of-the-art low-cost air quality sensors, assemblies, calibration and evaluation for respiration-associated diseases: A systematic review," *Atmosphere*, vol. 15, no. 4, p. 471, 2024.

[4] C. Malings et al., "Air quality estimation and forecasting via data fusion with uncertainty quantification: Theoretical framework and preliminary results," *Journal of Geophysical Research: Machine Learning and Computation*, vol. 1, no. 4, e2024JH000183, 2024.

[5] C. Zhang et al., "Unleashing the potential of geostationary satellite observations in air quality forecasting through artificial intelligence techniques," *Atmospheric Chemistry and Physics*, vol. 25, no. 2, pp. 759–770, 2025.

[6] B.-E. B. Semlali, C. El Amrani, G. Ortiz, J. Boubeta-Puig, and A. Garcia-de-Prado, "Sat-cep-monitor: An air quality monitoring software architecture combining complex event processing with satellite remote sensing," *Computers & Electrical Engineering*, vol. 93, p. 107 257, 2021.

[7] I. Vajs, D. Drajic, and Z. Cica, "Data-driven machine learning calibration propagation in a hybrid sensor network for air quality monitoring," *Sensors*, vol. 23, no. 5, p. 2815, 2023.

[8] N. Ajnoti, H. Gehlot, and S. N. Tripathi, "Hybrid instrument network optimization for air quality monitoring," *Atmospheric Measurement Techniques*, vol. 17, no. 6, pp. 1651–1664, 2024.

[9] Q. Liu, B. Cui, and Z. Liu, "Air quality class prediction using machine learning methods based on monitoring data and secondary modeling," *Atmosphere*, vol. 15, no. 5, p. 553, 2024.

[10] S. Saleh, A. Abohamama, and A. Tolba, "Intelligent real-time air quality index classification for smart home digital twins.," *International Journal of Advanced Computer Science & Applications*, vol. 16, no. 3, 2025.

[11] F. Mehmood, S. U. Rehman, and A. Choi, "Vision-aq: Explainable multimodal deep learning for air pollution classification in smart cities," *Mathematics*, vol. 13, no. 18, p. 3017, 2025.

[12] M. M. Rahman et al., "Airnet: Predictive machine learning model for air quality forecasting using web interface," *Environmental Systems Research*, vol. 13, no. 1, p. 44, 2024.

[13] M. A. Haq, "Smotednn: A novel model for air pollution forecasting and aqi classification.," *Computers, Materials & Continua*, vol. 71, no. 1, 2022.

[14] S. Kutala, H. Awari, S. Velu, A. Anthonisamy, N. J. Bathula, and S. Inthiyaz, "Hybrid deep learning-based air pollution prediction and index classification using an optimization algorithm.," *AIMS Environmental Science*, vol. 11, no. 4, 2024.

[15] M. W. Jalali, B. Saidi, H. Farahmand, M. A. R. Panah, and E. N. Saruhan, "Scalable ai-driven air quality forecasting and classification for public health applications," *Discover Atmosphere*, vol. 3, no. 1, p. 25, 2025.

[16] B. V. Jayadi, M. D. Lauro, Z. Rusdi, and T. Handhayani, "Air quality index classification for imbalanced data using machine learning approach," *Sistemasi: Jurnal Sistem Informasi*, vol. 13, no. 3, pp. 951–958, 2024.

[17] N. G. Rezk, S. Alshathri, A. Sayed, E. El-Din Hemdan, and H. El-Behery, "Sustainable air quality detection using sequential forward selection-based ml algorithms," *Sustainability*, vol. 16, no. 24, p. 10 835, 2024.

[18] Z. Xu, H. Zhang, A. Zhai, C. Kong, and J. Zhang, "Stacking ensemble learning and shap-based insights for urban air quality forecasting: Evidence from shenyang and global implications," *Atmosphere*, vol. 16, no. 7, p. 776, 2025.

[19] E. Kalantari, H. Gholami, H. Malakooti, A. R. Nafarzadegan, and V. Moosavi, "Machine learning for air quality index (aqi) forecasting: Shallow learning or deep learning?" *Environmental Science and Pollution Research*, vol. 31, no. 54, pp. 62 962–62 982, 2024.

[20] M. Maciejewska, A. Azizah, and A. Szczurek, "Iaq prediction in apartments using machine learning techniques and sensor data," *Applied Sciences*, vol. 14, no. 10, p. 4249, 2024. DOI: 10.3390/app14104249.

[21] S. Kumari and S. K. Singh, "Machine learning-based time-series models for effective co2 emission prediction in india," *International Journal of Sustainable Computing*, vol. 3, no. 2, pp. 67–78, 2022.

[22] A. Yenkikar, V. P. Mishra, M. Bali, and T. Ara, "Explainable forecasting of air quality index using a hybrid random forest and arima model," *MethodsX*, p. 103 517, 2025.

[23] C. Andenna and R. V. Gagliardi, "Insights into air quality index (aqi) variability with explainable machine learning techniques," *Environmental and Earth Sciences Proceedings*, vol. 34, no. 1, p. 1, 2025.

[24] A. S. Iffadah, D. A. Prasetya, et al., "Shapley additive explanations interpretation of the xgboost model in predicting air quality in jakarta," *Jurnal Riset Informatika*, vol. 7, no. 3, pp. 119–127, 2025.

[25] M. Rajesh, R. G. Babu, U. Moorthy, and S. V. Easwaramoorthy, "Machine learningdriven framework for realtime air quality assessment and predictive environmental health risk mapping," *Scientific Reports*, vol. 15, no. 1, p. 28 801, 2025.

[26] A. Das and H. Gupta, "Air quality prediction model for monitoring aqi," in *EPJ Web of Conferences*, EDP Sciences, vol. 328, 2025, p. 01 070.

[27] S. V. Chandrashekar et al., "A hybrid physics-informed neural and explainable ai approach for scalable and interpretable aqi predictions," *MethodsX*, p. 103 597, 2025.

[28] J. Wu et al., "A novel framework for high resolution air quality index prediction with interpretable artificial intelligence and uncertainties estimation," *Journal of Environmental Management*, vol. 357, p. 120 785, 2024.

[29] O. Ramírez, B. Hernández-Cuellar, and J. D. de la Rosa, "Air quality monitoring on university campuses as a crucial component to move toward sustainable campuses: An overview," *Urban Climate*, vol. 52, p. 101 694, 2023.

[30] M. Kozłowski, A. Asenov, V. Pencheva, S. A. Bęczkowska, A. Czerepicki, and Z. Zysk, "Autonomous system for air quality monitoring on the campus of the university of ruse: Implementation and statistical analysis," *Sustainability*, vol. 17, no. 14, p. 6260, 2025.

[31] N. Yadav et al., "Using deep transfer learning and satellite imagery to estimate urban air quality in data-poor regions," *Environmental Pollution*, vol. 342, p. 122 914, 2024.

[32] Y. Yang, Z. Hu, K. Bian, and L. Song, "Imgsensingnet: Uav vision guided aerial-ground air quality sensing system," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, IEEE, 2019, pp. 1207–1215.

[33] F. A. Panaite, C. Rus, M. Leba, A. C. Ionica, and M. Windisch, "Enhancing air-quality predictions on university campuses: A machine-learning approach to pm2.5 forecasting at the university of petroșani," *Sustainability*, vol. 16, no. 17, p. 7854, 2024.

[34] S. H. Godasiaei, O. A. Ejohwomu, H. Zhong, and D. Booker, "Integrating experimental analysis and machine learning for enhancing energy efficiency and indoor air quality in educational buildings," *Building and Environment*, vol. 276, p. 112 874, 2025.

[35] K. Sridhar, P. Radhakrishnan, G. Swapna, R. Kesavamoorthy, L. Pallavi, and R. Thiagarajan, "A modular iot sensing platform using hybrid learning ability for air quality prediction," *Measurement: Sensors*, vol. 25, p. 100 609, 2023.