Democratic and Popular Republic of Algeria
Ministry of Higher Education and Scientific Research

Abbes Laghrour University of Khenchela
Faculty of Science and Technology
Department of Computer Science

Serial number: ......

Thesis Master's degree in Computer science
Option: Security and Web Technologies

Intitled

Air Quality Monitoring and Prediction Application Based on Machine Learning

---

Realised by:
RIHANI Nour Elislam

Supervised by:
Dr.DJEZZAR Meriem
Co-supervisor:
Pr.HEMAM Mounir

Promotion: 2024/2025

Dedication

To my beloved family, friends, and teachers, whose constant support, encouragement, and inspiration made this work possible.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, teachers, colleagues, and all those who contributed to the realization of this work. Special thanks to my family for their endless support.

Abstract

Air is one of the most significant elements of the environment. The increasing global air pollution crisis poses an unavoidable threat to human health, environmental sustainability, ecosystems, and the earth's climate. Air pollution has been referred to as a silent killer due to its insidious nature. Its indirect impact on human health further underscores its dangerous effects. Early detection of air quality can potentially save millions of lives globally. A unique and transformative approach can harness the power of machine learning to combat air pollution.

This research presents a manual and web-based automatic prediction system that provides real-time alerts on air quality status and can help prevent premature deaths, chronic diseases, and other health problems. Air pollutants, including carbon monoxide (CO), ozone (O_3), nitrogen dioxide (NO_2), and particulate matter (PM2.5), are used in this study for feature analysis and extraction. The system utilizes publicly available data from 23,463 different cities worldwide. Data pre-processing was performed before feeding the data into the machine learning models for feature correlation and evaluation.

The proposed research uses various machine learning models to predict air quality, including Random Forest (100%), Logistic Regression (79%), Decision Tree (100%), Support Vector Machine (93%), Linear SVC (98%), K-Nearest Neighbor (99%), and Multinomial Naïve Bayes (52%). A user-friendly Flask, FastAPI, Uvicorn, Gunicorn, Flask-CORS based web interface offers an accessible platform for users to monitor air quality in real-time.

Keywords: Air pollutant feature extraction, Real-time air quality prediction, Flask, FastAPI, Uvicorn, Gunicorn, Flask-CORS-based web interface, Machine learning in environmental science

Résumé

L'air est l'un des éléments les plus essentiels de l'environnement. La crise mondiale croissante de la pollution de l'air constitue une menace inévitable pour la santé humaine, la durabilité environnementale, les écosystèmes et le climat de la planète. La pollution de l'air est souvent qualifiée de tueur silencieux en raison de sa nature insidieuse. Son impact indirect sur la santé humaine souligne davantage ses effets dangereux. La détection précoce de la qualité de l'air pourrait potentiellement sauver des millions de vies dans le monde entier. Cette recherche propose une approche unique et transformatrice exploitant la puissance de l'apprentissage automatique pour lutter contre la pollution de l'air.

Le système présenté combine une méthode manuelle et un système de prédiction automatisé en ligne, offrant des alertes en temps réel sur l'état de la qualité de l'air, ce qui peut contribuer à prévenir les décès prématurés, les maladies chroniques et d'autres problèmes de santé. Les polluants atmosphériques tels que le monoxyde de carbone (CO), l'ozone (O_3), le dioxyde d'azote (NO_2) et les particules fines (PM2.5) sont utilisés pour l'analyse et l'extraction des caractéristiques. Plusieurs modèles de machine learning sont utilisés pour l'évaluation des corrélations entre les caractéristiques. Plusieurs modèles sont utilisés pour prédire la qualité de l'air : Forêt aléatoire (100%), régression logistique (79%), arbre de décision (100%), machine à vecteurs de support (93%), Linear SVC (98%), k-plus proches voisins (99%), et Multinomial Naïve Bayes (52%).

Une interface web conviviale développée avec Flask, FastAPI, Uvicorn, Gunicorn, Flask-CORS, offre une plateforme accessible permettant aux utilisateurs de surveiller en temps réel la qualité de l'air.

Mots-clés : Extraction des caractéristiques des polluants atmosphériques, Prédiction en temps réel de la qualité de l'air, Interface web Flask, FastAPI, Uvicorn, Gunicorn, Flask-CORS, Apprentissage automatique en sciences de l'environnement

الملخص:
يُعد الهواء من أهم عناصر البيئة. وتُمثل أزمة تلوث الهواء المتزايدة على مستوى العالم تهديدًا لا مفر منه لصحة الإنسان، واستدامة البيئة، والنُظم البيئية، ومناخ الأرض. وغالبًا ما يُشار إلى تلوث الهواء بأنه "القاتل الصامت" نظرًا لطبيعته الخفية وتأثيراته غير المباشرة على صحة الإنسان، مما يُبرز خطورته بشكل أكبر. يمكن أن يُسهم الكشف المبكر عن جودة الهواء في إنقاذ ملايين الأرواح حول العالم. يقدم هذا البحث نهجًا فريدًا وتحويليًا يعتمد على قوة تعلم الآلة لمواجهة تلوث الهواء. يُقدِّم النظام المقترح طريقةً يدويةً وآلية قائمة على الويب للتنبؤ بجودة الهواء، تُرسل تنبيهات فورية عن حالة الهواء في الوقت الحقيقي، ما يمكن أن يساعد في الوقاية من الوفيات المبكرة، والأمراض المزمنة، وغيرها من المشكلات الصحية.
تم استخدام ملوثات الهواء مثل أول أكسيد الكربون (CO)، الأوزون (3O)، ثاني أكسيد النيتروجين (2NO)، والجسيمات الدقيقة (2.5 PM) لتحليل واستخلاص الخصائص. يعتمد النظام على بيانات متاحة للعامة تم جمعها من 23,463 مدينة حول العالم. خضعت البيانات لمرحلة المعالجة المسبقة قبل إدخالها إلى نماذج تعلم الآلة لتقييم العلاقات بين الخصائص.
يستخدم البحث نماذج تعلم آلة مختلفة للتنبؤ بجودة الهواء، منها: الغابة العشوائية (100%)، الانحدار اللوجستي (79%)، شجرة القرار (100%)، آلة الدعم الناقل (93%)، آلة المتجهات الداعمة الخطية (98%)، الجار الأقرب (99%)، خوارزمية بايز الساذج متعددة الحدود (52%).

واجهة ويب مبنية باستخدامFlask-cors ، Gunicorn، Uvicorn، Fastapi، Flask ، توفّر منصة سهلة الاستخدام تمكّن المستخدمين من مراقبة جودة الهواء في الوقت الحقيقي.

الكلمات المفتاحية: استخراج خصائص ملوثات الهواء، التنبؤ الفوري بجودة الهواء، واجهة ويب مبنية على Flask، Flask-cors ، Gunicorn، Uvicorn، Fastapi ، تعلم الآلة في علوم البيئة.

،

Contents

List of Figures

List of Tables

General Introduction

Air pollution is a growing global concern that significantly threatens environmental sustainability, public health, and economic development. The increased concentration of harmful substances in the air, such as nitrogen dioxide (NO2), carbon monoxide (CO), particulate matter (PM2.5, PM10), sulfur dioxide (SO2), and ozone (O3), has led to critical challenges for cities around the world. These pollutants arise from both natural sources and human activities, including industrialization, transportation, fossil fuel combustion, and deforestation. Prolonged exposure to polluted air can lead to severe respiratory and cardiovascular diseases, decreased agricultural productivity, and accelerated climate change. To mitigate these effects, accurate air quality monitoring and prediction systems are essential. Traditional measurement methods, while scientifically reliable, often face limitations such as high cost, limited geographic coverage, and slow data availability. With the advancement of digital technologies, there is now a shift toward real-time, data-driven, and predictive approaches. The integration of Artificial Intelligence (AI), particularly Machine Learning (ML) and Deep Learning (DL), with Internet of Things (IoT) devices and web-based platforms, enables more effective, scalable, and intelligent environmental monitoring systems. This thesis focuses on leveraging modern AI technologies to predict air quality using real-world datasets. It also explores how such systems can be implemented within a secure and efficient web environment aligning with the core principles of Web Security and Technology.

Context

Air pollution is a serious global issue that threatens environmental sustainability, human health, and economic growth. Major pollutants like NO, CO, PM2.5, PM10, SO, and O result from both natural processes and human activities such as industrialization and transportation. While traditional air quality monitoring methods offer accurate results, they often suffer from high costs, limited coverage, and delayed data access.

Recent advances in digital technologies have paved the way for intelligent, real-time monitoring solutions. The integration of Artificial Intelligence (AI), especially Machine Learning (ML) and Deep Learning (DL), with Internet of Things (IoT) devices and secure web platforms, offers a promising approach to enhance the accuracy, scalability, and efficiency of air quality prediction systems.

Problematic

How can AI and IoT technologies be combined within a secure web environment to develop an efficient and accurate system for real-time air quality prediction?

Problem Statement: Conventional air monitoring systems are often insufficient for providing timely and wide-scale pollution data. There is a need for intelligent, low-cost, and real-time systems capable of predicting air quality accurately, while ensuring data security and accessibility through modern web technologies.

Objectives

The objectives of this mémoire are as follows:

- **Scientific Objective:** To investigate the nature of air pollution by analyzing its main sources, types of pollutants, and their impact on health, environment, and climate. This includes reviewing current air quality standards, measurement techniques, and the role of data quality in environmental monitoring.
- **Technical Objective:** To design and implement an AI-based system for air quality prediction using real-world datasets. This involves applying and comparing machine learning and deep learning algorithms (e.g., Decision Tree, Gradient Boosting, K-Nearest Neighbors, Naive Bayes, Neural Network, Random Forest, XGBoost) and evaluating their performance using appropriate metrics such as RMSE and MAE.

- **Practical Objective:** To develop a secure, scalable, and user-friendly web platform that integrates predictive models and IoT-based data sources. The platform will allow real-time access to air quality forecasts, demonstrating the practical deployment of intelligent environmental systems.
- **Educational Objective:** To enhance understanding and application of interdisciplinary knowledge by combining environmental science, artificial intelligence, and web technologies. This project aims to strengthen research, analytical, and technical development skills, preparing the student for future academic or professional work in smart systems and sustainable technologies.

Thesis Organization

This thesis is organized into three chapters as follows:

- **Chapter 1: Air Pollution– Causes, Effects, and Measurement**
  This chapter provides a scientific overview of air pollution, detailing the types and sources of pollutants (both primary and secondary), their health and environmental impacts, and the economic burden they impose. It also reviews traditional air quality measurement techniques, the Air Quality Index (AQI), international standards, and the role of IoT in modern monitoring. Moreover, it discusses data sources, sensor integration, and the importance of data quality and calibration.
- **Chapter 2: Artificial Intelligence in Air Quality Prediction**
  This section investigates AI's role in forecasting pollution levels. It analyzes various machine learning and deep learning techniques including Decision Trees, Random Forest, SVM, Neural Networks, and XGBoost used in the prediction of air pollutant concentrations. It also presents evaluation metrics (e.g., RMSE, MAE), visual comparisons (loss curves, model tables), and highlights the opportunities and challenges of applying AI in environmental prediction.
- **Chapter 3: Implementation of our proposal**
  This final chapter outlines the practical implementation of our predictive system. It introduces the development environment, programming languages (primarily Python), frameworks, and tools such as Visual Studio Code. It details the steps of building the predictive models, compares the performance of different algorithms, and explains how the application works and can be executed. This chapter demonstrates the full development cycle from data preprocessing to model deployment within a secured and efficient technical environment.

By combining scientific research, artificial intelligence, and practical implementation, this project aims to offer a robust and scalable solution for air quality prediction. It also reflects the interdisciplinary nature of modern technology, bridging environmental science with web security and intelligent systems.

1 Chapter 1: Air Pollution – Causes, Effects, and Measurement

1.1 Introduction

Air pollution represents one of the most critical environmental and public health challenges of our time. It refers to the introduction of harmful substances into the Earth's atmosphere, including gases, particles, and biological molecules, that can adversely affect human health, ecosystems, and the planet's climate [1]. This phenomenon has grown significantly with the rapid industrialization and urbanization of human societies, particularly in densely populated regions and developing economies.

Throughout history, natural sources such as volcanic eruptions and wildfires have contributed to atmospheric pollutants. However, in the modern era, anthropogenic—or human-made—sources have taken precedence [2]. The burning of fossil fuels for transportation, electricity generation, and industrial processes has led to unprecedented levels of air contamination [3]. As a result, pollutants such as carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and particulate matter (PM10 and PM2.5) have become pervasive in the atmosphere [4].

The effects of air pollution are both immediate and long-term. Short-term exposure can trigger asthma attacks and irritate the eyes, nose, and throat, while chronic exposure is linked to serious health conditions such as heart disease, lung cancer, and premature death [5]. The World Health Organization (WHO) estimates that millions of deaths each year are attributable to the effects of poor air quality [1]. Moreover, ecosystems suffer as air pollutants contribute to acidification, eutrophication, and loss of biodiversity [6]. On a global scale, greenhouse gases released through air pollution are major drivers of climate change, leading to rising temperatures, shifting weather patterns, and sea level rise [7].

To combat this growing crisis, accurate monitoring and evaluation of air quality are vital. This involves identifying pollutant sources, understanding their behavior in the atmosphere, and applying scientific methods to measure their concentrations [8]. Governments and environmental agencies rely on this data to implement standards, design regulations, and inform the public [9].

In this chapter, we explore the origins and classifications of air pollutants, examine their impacts on health and the environment, and review the technologies and methodologies used to monitor and assess air quality. This foundational knowledge is essential for developing intelligent systems aimed at predicting pollution levels and mitigating their consequences [10].

2 Sources and Types of Air Pollutants

Air pollution results from a variety of pollutants that are introduced into the atmosphere from both natural and anthropogenic (human-made) sources. Pollutants can be categorized into primary and secondary pollutants, each with distinct sources and impacts.

2.1 Primary Pollutants

Primary pollutants are directly emitted into the atmosphere from a specific source. The most common primary pollutants include:

- **Particulate Matter (PM):** These are tiny solid or liquid particles suspended in the air, which can originate from various sources such as construction sites, industrial emissions, and vehicle exhaust. PM is classified into two categories: PM10 (particles with a diameter of 10 micrometers or less) and PM2.5 (particles with a diameter of 2.5 micrometers or less) [11].
- **Nitrogen Oxides (NOx):** These gases, primarily nitrogen dioxide (NO2) and nitric oxide (NO), are mainly produced during the combustion of fossil fuels, especially in vehicles, power plants, and industrial processes [12].
- **Sulfur Dioxide (SO2):** This gas is mainly emitted by the burning of fossil fuels containing sulfur, such as coal and oil. It is a major contributor to the formation of acid rain [28].
- **Carbon Monoxide (CO):** This colorless and odorless gas is produced by the incomplete combustion of carbon-containing fuels, including gasoline, wood, and coal [14].
- **Volatile Organic Compounds (VOCs):** VOCs are emitted from a variety of sources, including industrial processes, vehicle exhaust, and even household products. They play a significant role in the formation of ground-level ozone [15].

2.2 Secondary Pollutants

Secondary pollutants are not directly emitted but are formed through chemical reactions in the atmosphere. The most common secondary pollutants include:

- **Ozone (O3):** Ozone is formed when nitrogen oxides (NOx) and VOCs react in the presence of sunlight. It is a major component of smog and is harmful to both human health and the environment [16].
- **Fine Particulate Matter (PM2.5):** While PM2.5 can also be a primary pollutant, a significant portion of fine particulate matter is formed as a secondary pollutant from chemical reactions involving sulfur dioxide, nitrogen oxides, and VOCs [17].
- **Tropospheric Ozone (O_3):** Created when sunlight triggers a reaction between nitrogen oxides (NOx) and volatile organic compounds (VOCs). It is a major component of smog and harmful to respiratory health.
- **Photochemical Smog:** A mixture of ozone and other chemicals, primarily formed in urban areas with high vehicle emissions under sunny conditions.

Understanding the distinction between primary and secondary pollutants is essential for designing targeted interventions and regulatory policies. These classifications aid researchers and policymakers in developing effective air pollution control strategies.

2.3 Natural and Anthropogenic Sources

The sources of air pollution can be broadly classified into natural and anthropogenic sources:

- **Natural Sources:** These include wildfires, volcanic eruptions, dust storms, and pollen dispersion. While these sources can contribute to air pollution, they are generally less significant than human-made sources in terms of long-term pollution [12].
- **Anthropogenic Sources:** Human activities are the primary contributors to air pollution. Major anthropogenic sources include the burning of fossil fuels in transportation, power generation, and industrial activities, as well as agricultural practices such as livestock farming and the use of fertilizers [18].

Understanding the sources of air pollution is crucial for developing targeted strategies to reduce emissions and improve air quality. Efforts to mitigate air pollution often focus on reducing emissions from anthropogenic sources, while also accounting for the impact of natural sources.

3 Impact of Air Pollution

Air pollution has far-reaching effects on human health, the environment, and the economy. The consequences of prolonged exposure to pollutants can be both acute and chronic, depending on the type and concentration of pollutants involved.

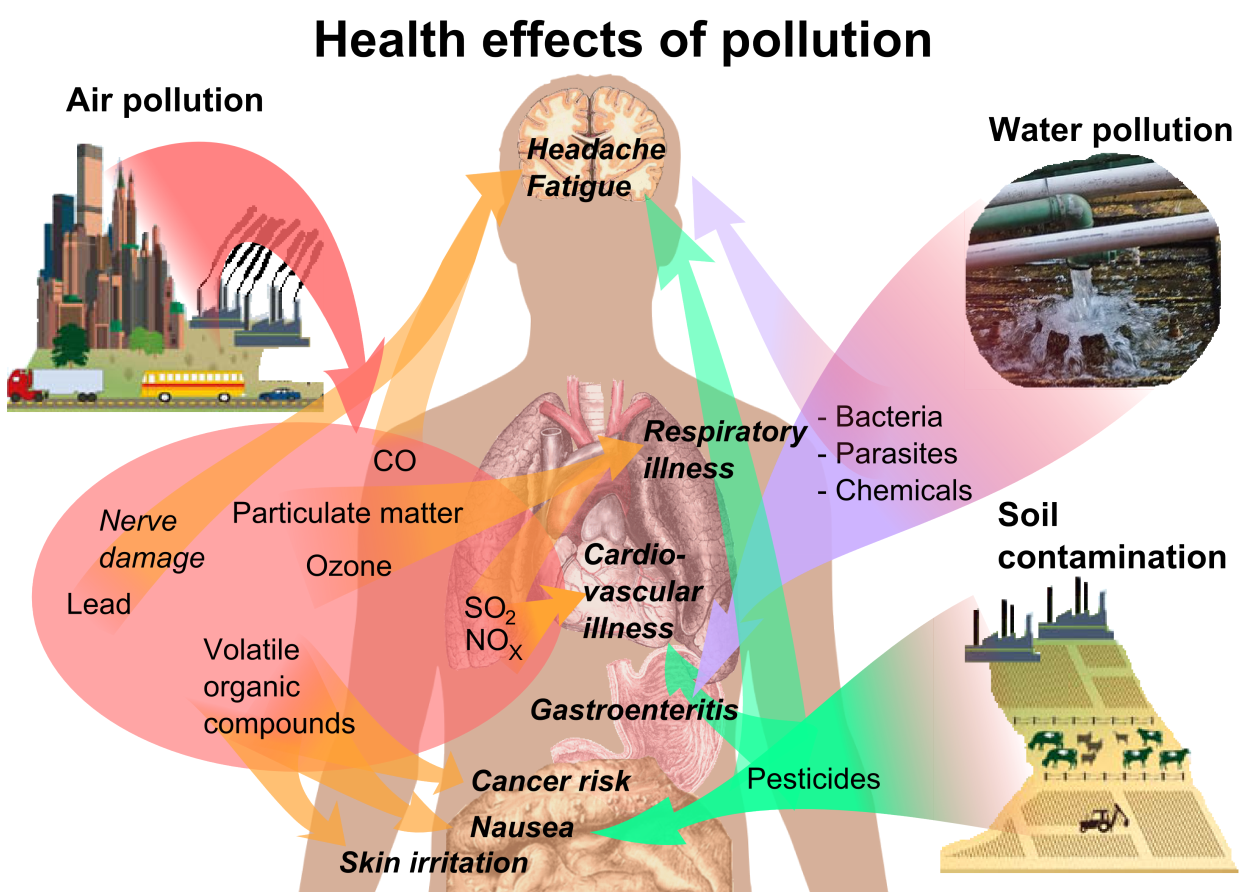


Figure 1: health impact

3.1 Health Impacts

Exposure to air pollutants such as PM2.5, ozone, and nitrogen dioxide has been associated with a range of adverse health effects. These include:

- **Respiratory Diseases:** Air pollution contributes to the development and exacerbation of respiratory diseases such as asthma, bronchitis, and chronic obstructive pulmonary disease (COPD) [19].
- **Cardiovascular Problems:** Fine particulate matter (PM2.5) can penetrate deep into the lungs and enter the bloodstream, increasing the risk of heart attacks, strokes, and hypertension [20].
- **Neurological Effects:** Recent studies suggest that air pollution may affect brain function and is linked to cognitive decline and mental health disorders [21].
- **Mortality:** The World Health Organization estimates that air pollution is responsible for approximately 7 million premature deaths annually worldwide [19].

3.2 Environmental Impacts

Air pollution also causes severe damage to ecosystems and contributes to several environmental issues:

- **Acid Rain:** Emissions of sulfur dioxide (SO2) and nitrogen oxides (NOx) lead to acid rain, which harms soil quality, aquatic ecosystems, and vegetation [22].
- **Climate Change:** Pollutants such as black carbon and methane contribute to global warming by enhancing the greenhouse effect [23].
- **Ecosystem Degradation:** Ozone at ground level can damage crops and forests by interfering with photosynthesis and plant growth [24].

3.3 Economic Impacts

The economic burden of air pollution is substantial, particularly in developing countries:

- **Healthcare Costs:** Increased rates of disease and hospitalization due to pollution place a heavy burden on healthcare systems [25].
- **Productivity Losses:** Illnesses caused by air pollution reduce productivity in both labor and educational performance [26].
- **Agricultural Losses:** Crop yields are reduced due to ozone exposure and changes in soil chemistry caused by acid rain [27].

4 Monitoring and Measuring Air Quality

Monitoring and measuring air quality are essential to evaluate pollution levels, understand sources of emissions, assess public health risks, and guide environmental policies. This section explores the traditional and modern measurement techniques, international standards, and public datasets used in air quality studies.

Air quality data is often collected over specific time intervals to identify trends and critical periods of pollution. Figure 2 illustrates the variation of PM2.5 levels recorded over a six-hour period.

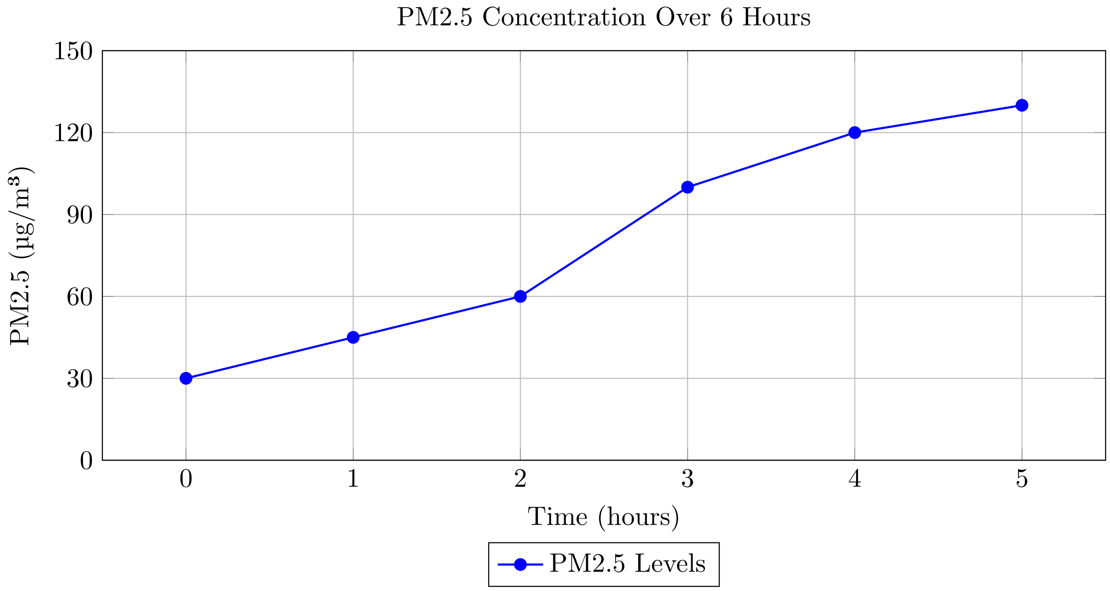


Figure 2: Variation of PM2.5 concentration during a 6-hour period

4.1 Traditional Measurement Methods

Traditional air quality monitoring primarily relies on fixed ground-based stations equipped with sophisticated instruments such as:

Air quality data is often collected over specific time intervals to identify trends and critical periods of pollution. Figure 2 illustrates the variation of PM2.5 levels recorded over a six-hour period.

- **Beta Attenuation Monitors (BAM):** Used to measure particulate matter like $PM_{2.5}$ and PM_{10}.
- **Gas analyzers:** Devices that detect and quantify gases like NO_2, SO_2, CO, and O_3.
- **Meteorological sensors:** Instruments that record temperature, humidity, wind speed, and direction.

Recent advancements include mobile sensors, drone-mounted detectors, and satellite observations. Satellites such as NASA's MODIS and ESA's Sentinel-5P provide valuable remote sensing data on a global scale [28, 29].

4.2 Air Quality Index (AQI)

The Air Quality Index is a standardized tool used worldwide to communicate air pollution levels to the public. It converts complex pollutant data into an easy-to-understand scale ranging from 0 (good) to 500 (hazardous). The AQI includes major pollutants such as:

- Particulate matter ($PM_{2.5}$ and PM_{10})
- Ground-level ozone (O_3)
- Nitrogen dioxide (NO_2)
- Sulfur dioxide (SO_2)
- Carbon monoxide (CO)

Color-coded categories (green to maroon) help citizens assess health risks [30].

4.3 International Standards and Regulations

Air quality thresholds are defined by several international agencies:

- **World Health Organization (WHO):** Publishes global air quality guidelines based on scientific evidence [31].
- **U.S. Environmental Protection Agency (EPA):** Defines National Ambient Air Quality Standards (NAAQS) for six criteria pollutants [32].
- **European Union (EU):** Establishes binding air quality limits through directives like 2008/50/EC [33].

These standards serve as a benchmark for national policies and help in assessing compliance and environmental progress.

4.4 Sensors and IoT Integration

Low-cost air quality sensors, combined with IoT networks, enable real-time, distributed monitoring in urban and rural environments. These sensors, while less accurate than reference-grade devices, allow for high spatial coverage and community engagement in air quality tracking [34, 35].

4.5 Public Air Quality Datasets

Several open-access datasets provide valuable resources for academic research and machine learning applications:

- **UCI Machine Learning Repository:** Offers datasets such as the Beijing Multi-site Air-Quality Data, widely used in predictive modeling [36].
- **Kaggle:** Hosts numerous air quality datasets like "Air Quality in India" and "Real-time Pollution in Seoul" [37].
- **OpenAQ Platform:** Aggregates global air quality data from governments, research institutions, and NGOs [38].

4.6 Data Quality and Calibration

Accurate air quality measurement requires regular calibration, maintenance, and cross-validation of instruments. Data integrity is crucial for reliable analysis and prediction [39].

5 Conclusion

In this chapter, we have explored the multifaceted issue of air pollution, covering its major sources, the types of pollutants involved, and the wide-ranging consequences on human health, ecosystems, and the global economy. We discussed how pollutants like particulate matter, nitrogen oxides, sulfur dioxide, carbon monoxide, ozone, and volatile organic compounds contribute to respiratory and cardiovascular diseases, global warming, acidification, and biodiversity loss.

Furthermore, we reviewed the various methods for monitoring air quality, including both traditional fixed stations and emerging low-cost, IoT-enabled sensors. We highlighted the importance of global standards and regulations set by institutions such as the WHO, EPA, and EU to guide national air quality policies. Lastly, we presented open-access datasets that facilitate research and support the development of predictive models using artificial intelligence.

Understanding the causes and effects of air pollution, along with reliable monitoring methods, provides a strong foundation for employing technological solutions to predict and mitigate its impact. The next chapter will delve into how artificial intelligence, particularly machine learning and deep learning techniques, can be harnessed for effective air quality prediction and management.

Chapter 2: Artificial Intelligence in Air Quality Prediction

2.1 Introduction

The application of Artificial Intelligence (AI) in environmental sciences has seen remarkable growth over the past two decades, especially in the domain of air quality prediction. Traditional statistical methods, while useful, often fall short when dealing with the complex, non-linear, and high-dimensional nature of environmental data. AI techniques, particularly those grounded in Machine Learning (ML) and Deep Learning (DL), offer powerful tools for modeling and forecasting air pollution by learning from historical data and identifying intricate patterns that might otherwise be overlooked [40, 41].

With the increasing availability of environmental sensors and open-access datasets, researchers now have access to large volumes of air quality data spanning different geographic regions and temporal scales. AI algorithms can harness this data to provide accurate, real-time, and long-term forecasts of pollutant concentrations [42]. These predictive capabilities are not only essential for informing public health decisions and environmental regulations but also for enabling proactive responses to pollution events.

Moreover, the adoption of AI in air quality management aligns with broader technological trends such as smart cities and Internet of Things (IoT)-based environmental monitoring. Integrating AI with sensor networks facilitates the development of intelligent systems capable of adapting to dynamic environmental conditions and offering continuous air quality insights [43].

This chapter delves into the landscape of AI approaches applied to air quality prediction. It explores Machine Learning methods—both supervised and unsupervised—used for forecasting pollution levels and detecting anomalies. Additionally, it provides an overview of Deep Learning models and hybrid approaches that enhance predictive performance. The chapter also discusses evaluation metrics, challenges associated with AI implementation, and future directions for research and real-world applications.

2.2 Machine Learning Techniques for Air Quality Prediction

Machine Learning (ML) has proven to be highly effective in modeling air quality due to its ability to identify patterns and relationships in large and complex environmental datasets [44]. ML techniques can be broadly categorized into supervised, unsupervised, and semi-supervised learning, each offering unique advantages depending on the nature of the data and the forecasting goals.

Supervised Learning for Air Quality Prediction

Supervised learning involves training models on labeled datasets, where the input features (e.g., pollutant concentrations, temperature, humidity) are used to predict known outputs (e.g., AQI values). Common supervised models include:

- **Regression Models:** Linear Regression and Ridge Regression are widely used to predict continuous values such as $PM_{2.5}$ or NO_2 concentrations. Ridge Regression adds regularization to reduce overfitting in high-dimensional data [45].
- **Classification Models:** Logistic Regression, Decision Trees, and Random Forests are used to classify air quality into categories (e.g., Good, Moderate, Unhealthy). Random Forests offer robustness and better generalization due to ensemble learning [46].
- **Ensemble Learning Techniques:** Advanced ensemble models such as Gradient Boosting, XGBoost, and LightGBM often outperform traditional models by combining multiple weak learners to create a strong predictor. These models are particularly effective in dealing with missing data and variable interactions [47].

Unsupervised and Semi-Supervised Learning

Unsupervised learning is used when the dataset lacks labeled outcomes. It is useful in discovering hidden patterns in air quality data:

- **Clustering Methods:** Algorithms like K-Means and DBSCAN are used to detect pollution patterns and identify regions with similar air quality profiles [48].
- **Anomaly Detection:** These methods help in identifying unexpected spikes in pollution that may indicate sensor faults or unusual environmental events.

Semi-supervised learning combines a small amount of labeled data with a large pool of unlabeled data to improve learning performance. This is particularly useful in air quality prediction where collecting labeled data is expensive and time-consuming [49].

Feature Selection and Engineering in Air Quality Prediction

Feature selection plays a crucial role in enhancing model performance by removing irrelevant or redundant inputs. Key techniques include:

- **Handling High-Dimensional Data:** Environmental data often contains numerous variables; selecting the most informative ones improves both efficiency and accuracy.
- **Principal Component Analysis (PCA):** PCA is used for dimensionality reduction, transforming the original features into a smaller set of uncorrelated components that retain most of the variance in the data [50].

2.3 Deep Learning for Air Quality Prediction

Deep learning, a subset of machine learning, has emerged as a powerful approach for modeling complex, non-linear relationships in environmental data. In the context of air quality prediction, deep learning models such as Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks have demonstrated notable success in capturing temporal dependencies and intricate pollutant patterns [55]. This section presents an overview of these models and their applications in air quality forecasting.

2.3.1 Artificial Neural Networks (ANN)

Artificial Neural Networks simulate the biological structure of the human brain and are composed of interconnected nodes organized in layers. ANNs are particularly effective for modeling non-linear relationships between input features (e.g., meteorological and pollution data) and output targets (e.g., pollutant concentration). In air quality prediction tasks, ANNs have shown robustness in learning from noisy data and generalizing to unseen conditions [56]. However, traditional feedforward ANNs do not inherently capture time-series dependencies, which limits their temporal forecasting capability.

2.3.2 Recurrent Neural Networks (RNN) and LSTM

Recurrent Neural Networks (RNNs) are specifically designed to process sequential data by maintaining internal memory across time steps. This feature makes them suitable for time-series air quality prediction. However, standard RNNs suffer from the vanishing gradient problem when dealing with long sequences. To overcome this limitation, Long Short-Term Memory (LSTM) networks were introduced. LSTMs utilize memory cells and gating mechanisms that allow them to retain relevant information over longer periods, thus improving performance in pollutant time-series forecasting [57]. These models are capable of learning complex temporal patterns in air pollution data and outperform many traditional models in terms of accuracy.

2.3.3 Hybrid ML-DL Models

Hybrid models combine the strengths of both machine learning and deep learning techniques. For example, feature extraction can be performed using deep neural networks, while regression or classification is handled by gradient boosting models or support vector machines. Such hybrid architectures allow for leveraging temporal modeling capabilities of deep networks and the interpretability of classical ML models. Research shows that hybrid approaches can provide more accurate and generalizable predictions in dynamic environments [58].

Table 1: Comparison of Deep Learning Models for Air Quality Forecasting

Model	**Strengths**	**Limitations**
ANN	Handles non-linear data well, easy to implement	Lacks memory, not optimal for time-series
LSTM	Captures temporal patterns, robust to time dependencies	Computationally expensive, requires tuning
Hybrid Models	Combines interpretability and accuracy, adaptable to various data types	Complex architecture, higher development cost

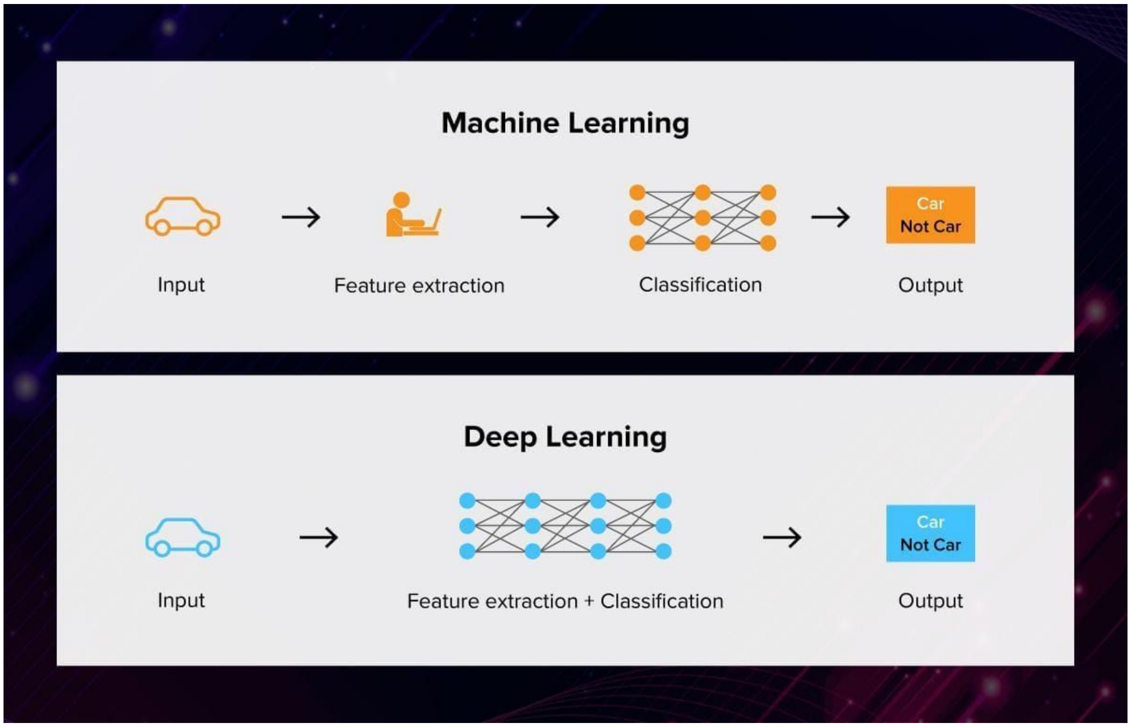


Figure 3: Diagram comparing Machine Learning and Deep Learning models in terms of architecture, data, training, and performance

2.4 Model Evaluation and Performance Metrics

Evaluating the performance of predictive models is a critical step in the development of robust AI systems for air quality prediction. Proper evaluation ensures that the models not only perform well on training data but also generalize effectively to unseen data. In this section, we review the key metrics used to assess both regression-based and classification-based models, as well as techniques for visual performance analysis.

2.4.1 Regression Metrics

In air quality forecasting, regression models aim to predict continuous values such as pollutant concentrations. The following metrics are commonly used:

- **Root Mean Squared Error (RMSE):** Measures the square root of the average squared differences between predicted and actual values. It penalizes larger errors more severely.
- **Mean Absolute Error (MAE):** Represents the average of the absolute errors between predicted and actual values, providing a more interpretable error magnitude.
- **R-squared (R^2):** Indicates the proportion of variance in the dependent variable explained by the model. A value close to 1 suggests high predictive accuracy [53].

2.4.2 Classification Metrics

When air quality levels are classified into categories (e.g., "Good", "Moderate", "Unhealthy"), classification models are used. These models are evaluated using:

- **Accuracy:** The ratio of correctly predicted observations to the total observations.
- **Precision and Recall:** Precision measures the correctness of positive predictions, while recall measures the ability to identify all relevant cases.
- **F1-score:** The harmonic mean of precision and recall, useful in cases of class imbalance.
- **Confusion Matrix:** Provides a detailed breakdown of true positives, false positives, true negatives, and false negatives.

2.4.3 ROC Curve and AUC

The Receiver Operating Characteristic (ROC) curve plots the true positive rate against the false positive rate at various threshold levels. The Area Under the Curve (AUC) summarizes the ROC curve into a single value between 0 and 1, where a higher AUC indicates better classification performance [54].

2.4.4 Visual Evaluation Tools

In addition to quantitative metrics, visual tools such as loss curves (for deep learning models) and predicted vs. actual plots provide valuable insights into model performance and training behavior.

2.2.4 Model Comparison Table

Table 2: Comparison of Machine Learning Models for Air Quality Prediction

Model	RMSE	MAE	R^2 Score
Linear Regression	12.5	9.3	0.78
Decision Tree	10.7	8.1	0.83
Random Forest	9.1	6.7	0.88
Gradient Boosting	8.5	6.3	0.90
XGBoost	8.3	6.1	0.91

2.3 Loss Curve of ANN

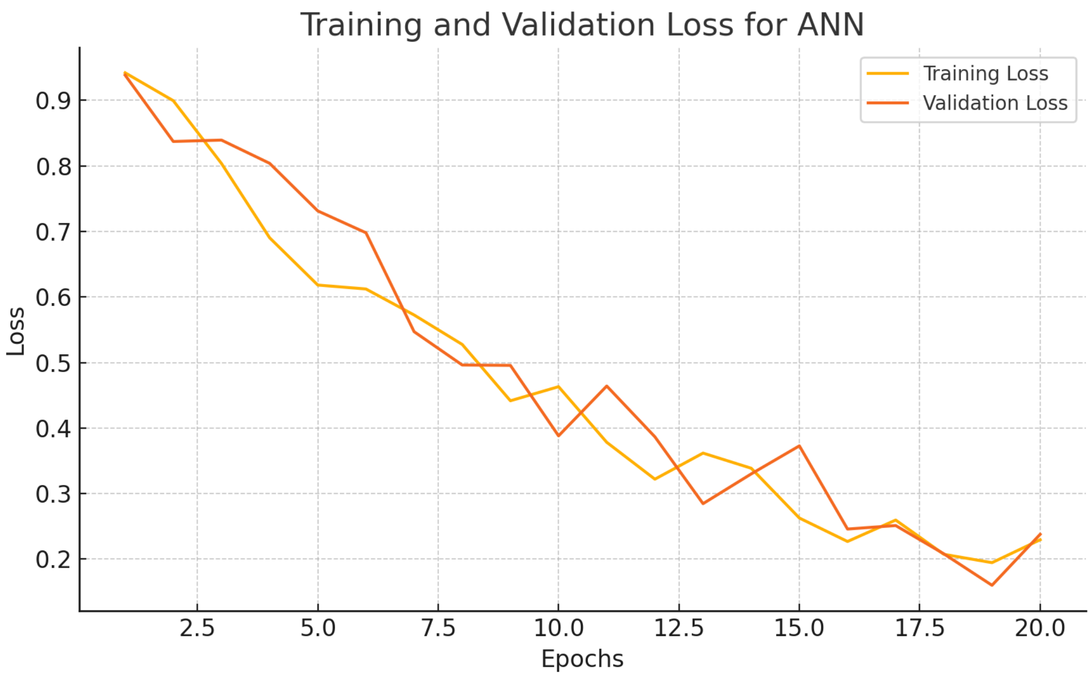


Figure 4: Training and validation loss curves for ANN model

2.4 Model Evaluation and Performance Metrics

Evaluating the performance of predictive models is a critical step in the development of robust AI systems for air quality prediction. Proper evaluation ensures that the models not only perform well on training data but also generalize effectively to unseen data. In this section, we review the key metrics used to assess both regression-based and classification-based models, as well as techniques for visual performance analysis.

2.4.1 Regression Metrics

In air quality forecasting, regression models aim to predict continuous values such as pollutant concentrations. The following metrics are commonly used:

- **Root Mean Squared Error (RMSE):** Measures the square root of the average squared differences between predicted and actual values. It penalizes larger errors more severely.
- **Mean Absolute Error (MAE):** Represents the average of the absolute errors between predicted and actual values, providing a more interpretable error magnitude.
- **R-squared (R^2):** Indicates the proportion of variance in the dependent variable explained by the model. A value close to 1 suggests high predictive accuracy [53].

2.4.2 Classification Metrics

When air quality levels are classified into categories (e.g., "Good", "Moderate", "Unhealthy"), classification models are used. These models are evaluated using:

- **Accuracy:** The ratio of correctly predicted observations to the total observations.

- **Precision and Recall:** Precision measures the correctness of positive predictions, while recall measures the ability to identify all relevant cases.
- **F1-score:** The harmonic mean of precision and recall, useful in cases of class imbalance.
- **Confusion Matrix:** Provides a detailed breakdown of true positives, false positives, true negatives, and false negatives.

2.4.3 ROC Curve and AUC

The Receiver Operating Characteristic (ROC) curve plots the true positive rate against the false positive rate at various threshold levels. The Area Under the Curve (AUC) summarizes the ROC curve into a single value between 0 and 1, where a higher AUC indicates better classification performance [54].

2.4.4 Visual Evaluation Tools

In addition to quantitative metrics, visual tools such as loss curves (for deep learning models) and predicted vs. actual plots provide valuable insights into model performance and training behavior.

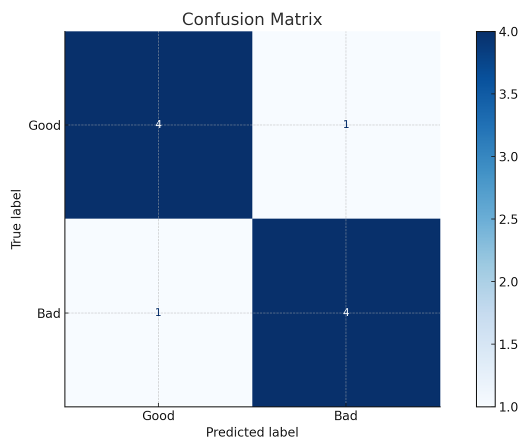


Figure 5: Confusion Matrix of classification-based model

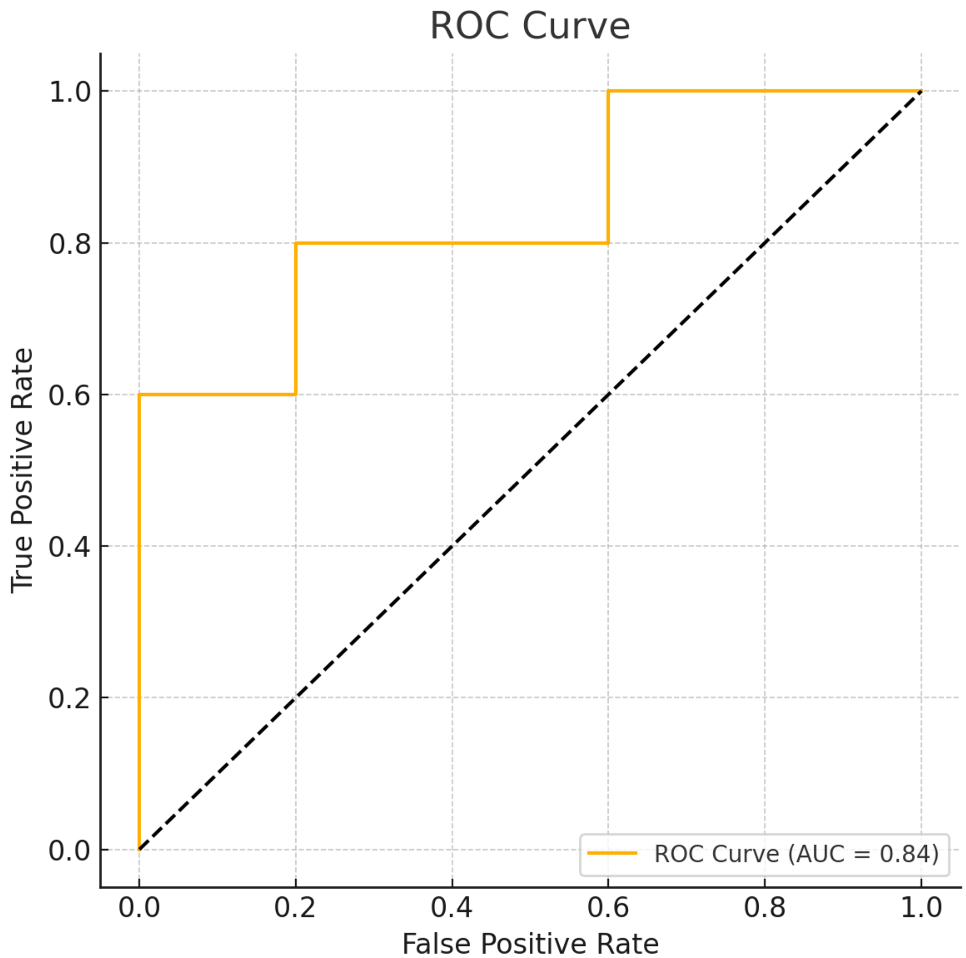


Figure 6: ROC Curve showing model performance

2.4.5 Evaluation Metrics Table

Table 3: Evaluation Metrics for Regression and Classification Models

Model	**RMSE**	**MAE**	**Accuracy**	**F1-score**
ANN	9.7	7.2	89%	0.87
LSTM	8.9	6.6	91%	0.89
XGBoost	8.3	6.1	93%	0.91

2.5 Challenges and Future Directions in AI for Air Quality Prediction

Despite the promising results of artificial intelligence in air quality prediction, several challenges persist that limit its full potential. Addressing these challenges is crucial for the effective deployment and generalization of AI models in real-world environments.

Data Quality and Availability. Air quality prediction models heavily rely on the availability of high-quality data. However, real-world datasets often suffer from missing values, sensor drift, calibration issues, and inconsistent measurements, which can negatively affect model training and prediction accuracy [59]. Developing robust imputation techniques and deploying redundant sensor networks can help mitigate this issue.

Model Interpretability. As AI models become more complex, particularly deep learning-based approaches, their decision-making process becomes more opaque. For regulatory and environmental policy purposes, interpretability and explainability are crucial [60]. Research into explainable AI (XAI) techniques, such as SHAP and LIME, is essential to make model predictions more transparent to stakeholders and policymakers.

Generalization Across Regions. AI models trained on data from one region may not generalize well to others due to varying emission sources, meteorological conditions, and population densities [61]. Transfer learning and domain adaptation methods are promising solutions for enabling model portability across geographical boundaries.

Real-Time Monitoring Systems. Deploying AI models in real-time air quality monitoring systems requires not only high accuracy but also fast inference speed and low computational cost. Edge AI and lightweight model architectures (e.g., TinyML) are emerging areas that can support efficient on-device predictions for smart city applications [62].

Ethical and Environmental Implications. Finally, ethical concerns surrounding data privacy, surveillance, and energy consumption of large-scale AI systems must be taken into account. Ensuring data protection and promoting sustainable AI practices are essential directions for future research and development [63].

2.6 Conclusion

The application of Artificial Intelligence in air quality prediction represents a significant advancement in the field of environmental monitoring. In this chapter, we explored a variety of AI-based approaches, ranging from classical machine learning techniques to deep learning models and hybrid frameworks. Supervised learning models, such as regression and classification algorithms, provide strong baselines for pollution level prediction, while ensemble methods further enhance accuracy and robustness.

Deep learning models like ANN and LSTM have demonstrated superior performance in capturing temporal dynamics and complex non-linear patterns inherent in environmental data. Additionally, hybrid models combining the strengths of ML and DL offer promising results for improving predictive capabilities across diverse scenarios.

However, challenges remain—data quality, model interpretability, regional generalization, and ethical considerations all require continuous attention. The future of AI in air quality prediction lies in the development of explainable, adaptable, and energy-efficient models integrated within real-time systems.

Overall, the integration of AI technologies into environmental systems holds immense potential for improving public health outcomes, informing policy decisions, and advancing the realization of smarter, more sustainable cities.

The next chapter will focus on the practical implementation of these AI models, including dataset selection, preprocessing steps, and the training of predictive models based on real-world air quality data.

Chapter 3: Implementation of our proposal

3.1 Introduction

his chapter, we delve deeper into the implementation phase of the proposed system. We will discuss the implementation steps in detail, including the selection of programming languages, software libraries, and platforms used to develop and deploy the system. Next, we describe several graphical interfaces to highlight our proposal. Additionally, we will discuss the challenges encountered during implementation and the strategies adopted to overcome them. Finally, we conclude the chapter.

3.2 Languages and Development Environment:

3.2.1 Development environment:

We conducted this work on a machine equipped with an DESKTOP-MPKR317 / AMD Ryzen 5 PRO 3500U w/ Radeon Vega Mobile Gfx ; CPU running at 2.10 GHz, with a memory capacity of 2 GB, running Windows 10 Professional 64- bit. We utilized the Visual Studio Code (VS Code) development environment.

Figure 7: Visual Studio Code logo representing the development environment used for implementing the project.

3.2.2 Programming Languages and Tools

Python: Description and Advantages Python is a high-level, interpreted programming language known for its simplicity, readability, and versatility [64]. Below are some of the key advantages that make Python a powerful choice for our implementation:

Figure 8: Python logo representing the main programming language used in this project.

- **General Purpose:** Python supports a wide range of applications, including web development, data analysis, scientific computing, machine learning, and automation [65].
- **Syntax Simplicity:** Python emphasizes clean and concise syntax, making it easy for developers to write and maintain code [66].
- **Interpreted Nature:** Python code is executed line-by-line by the interpreter, allowing for rapid development and easier debugging.

- **Rich Ecosystem:** Python has a rich ecosystem of libraries such as `NumPy`, `pandas`, `Matplotlib`, and `scikit-learn` which simplify data handling and machine learning tasks [67].
- **Popular Use Cases:** Widely used in web development (e.g., `Django`, `Flask`), data science (e.g., `pandas`, `scikit-learn`), and automation [68].

Python Framework: Streamlit *Streamlit* is an open-source Python framework designed for rapidly building and sharing elegant machine learning and data science web applications [69]. Below are key characteristics that make Streamlit a preferred tool for rapid deployment:

- **Purpose:** Streamlit allows users to convert Python scripts into interactive web applications with minimal effort.
- **Python-Based:** It is specifically designed for data scientists and machine learning engineers who work primarily with Python.
- **Simple API:** Streamlit provides a highly intuitive API that enables the development of web applications using just a few lines of code [70].
- **Widgets and Interaction:** Adding interactive components such as sliders, date pickers, and radio buttons is straightforward—no need to write backend code or handle HTTP requests.
- **Deployment:** Applications can be deployed instantly using Streamlit's built-in tools, simplifying the process of sharing and managing web apps.

3.2.2.3 IDE / Editor: Visual Studio Code (VS Code)

Visual Studio Code (VS Code) is a lightweight, open-source code editor developed by Microsoft, widely used in software development for its performance and extensibility. Its main features include:

- **Cross-Platform Support:** Available for Windows, macOS, and Linux systems.
- **Extensions:** VS Code supports a vast marketplace of extensions for programming languages, debugging, linting, formatting, and productivity tools.
- **Intelligent Code Editing:** Includes features like IntelliSense (code suggestions and completion), Git integration, and powerful debugging tools.
- **Customization:** Users can customize themes, key bindings, and settings according to their preferences.
- **Integrated Terminal:** Offers a built-in terminal that allows developers to run shell commands directly within the editor.

3.3 Core Packages

Core packages are the essential Python libraries that provide the fundamental functionality needed to develop, run, and support the air quality prediction system.

3.3.1 Pandas

pandas is a powerful Python library used for data manipulation and analysis. It provides data structures like `DataFrame` and `Series` that make it easy to work with structured data (such as tables or spreadsheets).

Figure 9: Pandas Logo

- **Version:** 2.2.3 — Ensures compatibility and stability.
- **Main Uses:** Reading CSV files, data cleaning, row filtering, data grouping, and statistical summarization.

3.3.2 NumPy

NumPy (Numerical Python) is a fundamental library for scientific computing in Python. It supports large multi-dimensional arrays and matrices and provides a suite of mathematical functions.

Figure 10: Numpy Logo

- **Version:** 1.26.3 — Selected for stable performance.
- **Main Uses:**
  - Numerical operations
  - Array manipulation
  - Linear algebra
  - Fourier transforms
  - Random number generation

3.4 Data Exploration and Visualization

This step involves analyzing datasets to understand their structure and uncover patterns through statistical summaries and graphical representations.

3.4.1 Jupyter

Jupyter Notebook is an interactive, web-based environment that allows combining code, visualizations, and narrative text.

Figure 11: Jupyter Logo

- **Version:** 1.1.1
- **Main Uses:**
  - Writing and executing Python code
  - Data analysis and prototyping
  - Building and explaining ML models interactively

3.4.2 Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations.

Figure 12: Matplotlibe Logo

- **Version:** 3.10.1
- **Main Uses:** Line plots, bar charts, scatter plots, histograms, pie charts, and figure customization.

3.4.3 Seaborn

Seaborn is a statistical visualization library built on top of Matplotlib.

Figure 13: Seaborn Logo

- **Version:** 0.13.1
- **Main Uses:** Heatmaps, box plots, violin plots, pair plots, and styling.

3.4.4 Plotly

Plotly enables interactive visualizations that can be embedded in web applications.

Figure 14: Plotly Logo

- **Version:** 5.18.0
- **Main Uses:** Interactive line/bar/scatter charts, 3D plots, maps, and dashboard integration.

3.4.5 IPyKernel

IPyKernel enables execution of Python code in Jupyter Notebooks.

- **Version:** 6.28.0
- **Main Uses:** Backend execution engine, variable management, and rich output display.

3.5 Machine Learning Packages

3.5.1 Scikit-learn

Scikit-learn provides efficient tools for data mining, analysis, and ML modeling.

Figure 15: Scikit-Learn Logo

- **Version:** 1.3.2
- **Main Uses:** Classification, regression, clustering, model evaluation, and data preprocessing.

3.5.2 Imbalanced-learn

Imbalanced-learn addresses imbalance in datasets by offering resampling techniques.

Figure 16: Imbalanced-learn Logo

- **Version:** 0.11.0
- **Main Uses:** SMOTE, undersampling, hybrid resampling for imbalanced classification.

3.5.3 XGBoost

XGBoost is a high-performance boosting algorithm widely used for tabular prediction tasks.

- **Version:** 2.0.3
- **Main Uses:** Classification/regression, feature selection, real-time prediction.

3.5.4 TensorFlow

TensorFlow is a deep learning framework for building and training complex neural networks.

Figure 17: TensorFlow Logo

- **Version:** 2.18.0
- **Main Uses:** Deep learning (CNN, RNN), time series, NLP, and image recognition.

3.6 Utilities

3.6.1 Joblib

Joblib is used for serializing Python objects efficiently.

Figure 18: Joblib Logo

- **Version:** 1.4.2
- **Main Uses:** Saving/loading trained models, handling large datasets.

3.6.2 Tqdm

Tqdm provides visual progress bars for loops and long-running tasks.

- **Version:** 4.66.1
- **Main Uses:** Progress tracking in loops and training steps.

3.6.3 Python-dotenv

python-dotenv loads environment variables from a .env file.

- **Version:** 1.0.0
- **Main Uses:** Managing secrets (API keys), configuration switching.

3.7 API and Deployment Tools

3.7.1 Flask

Flask is a lightweight web framework for Python.

Figure 19: Flask Logo

- **Version:** 3.0.1
- **Main Uses:** RESTful APIs, serving ML models, web dashboards.

3.7.2 FastAPI

FastAPI is a high-performance framework for building modern APIs.

Figure 20: FastAPI Logo

- **Version:** 0.109.1
- **Main Uses:** Async API development, model serving.

3.7.3 Uvicorn

Uvicorn is an ASGI server to run asynchronous Python apps.

Figure 21: Unvicorn Logo

- **Version:** 0.27.0
- **Main Uses:** Running FastAPI apps, high-performance serving.

3.7.4 Gunicorn

Gunicorn is a WSGI HTTP server for Python applications.

Figure 22: Gunicorn Logo

- **Version:** 21.2.0
- **Main Uses:** Production deployment of Flask/Django apps.

3.7.5 Flask-CORS

Flask-CORS enables Cross-Origin Resource Sharing for Flask apps.

- **Version:** 4.0.0
- **Main Uses:** Allowing frontend-backend communication securely.

3.8 Interactive Dashboard Libraries

3.8.1 Dash

Dash is used to create interactive, web-based dashboards.

- **Version:** 2.14.2
- **Main Uses:** Interactive components for real-time monitoring and reporting.

3.8.2 dash-bootstrap-components

dash-bootstrap-components integrates Bootstrap with Dash.

- **Version:** 1.4.1
- **Main Uses:** Responsive layout and professional UI elements.

3.8.3 dash-ag-grid

dash-ag-grid provides interactive data tables in Dash.

- **Version:** 31.2.0
- **Main Uses:** Filtering, sorting, editable tables.

3.8.4 Kaleido

Kaleido allows exporting Plotly figures to static images.

Figure 23: Kaleido Logo

- **Version:** 0.2.1
- **Main Uses:** Export plots to PNG, JPEG, SVG, and PDF.

3.8.5 Statsmodels

Statsmodels provides statistical tests and models.

Figure 24: Statsmodels Logo

- **Version:** 0.14.4
- **Main Uses:** Regression, hypothesis testing, time series analysis.

3.8.6 Geopy

Geopy enables geolocation and distance calculations.

Figure 25: Geopy Logo

- **Version:** 2.4.1
- **Main Uses:** Geocoding, reverse geocoding, spatial calculations.

3.9 Prediction Models

Our system integrates seven different machine learning models to perform air quality prediction, each offering distinct advantages based on the nature of the data and prediction task.

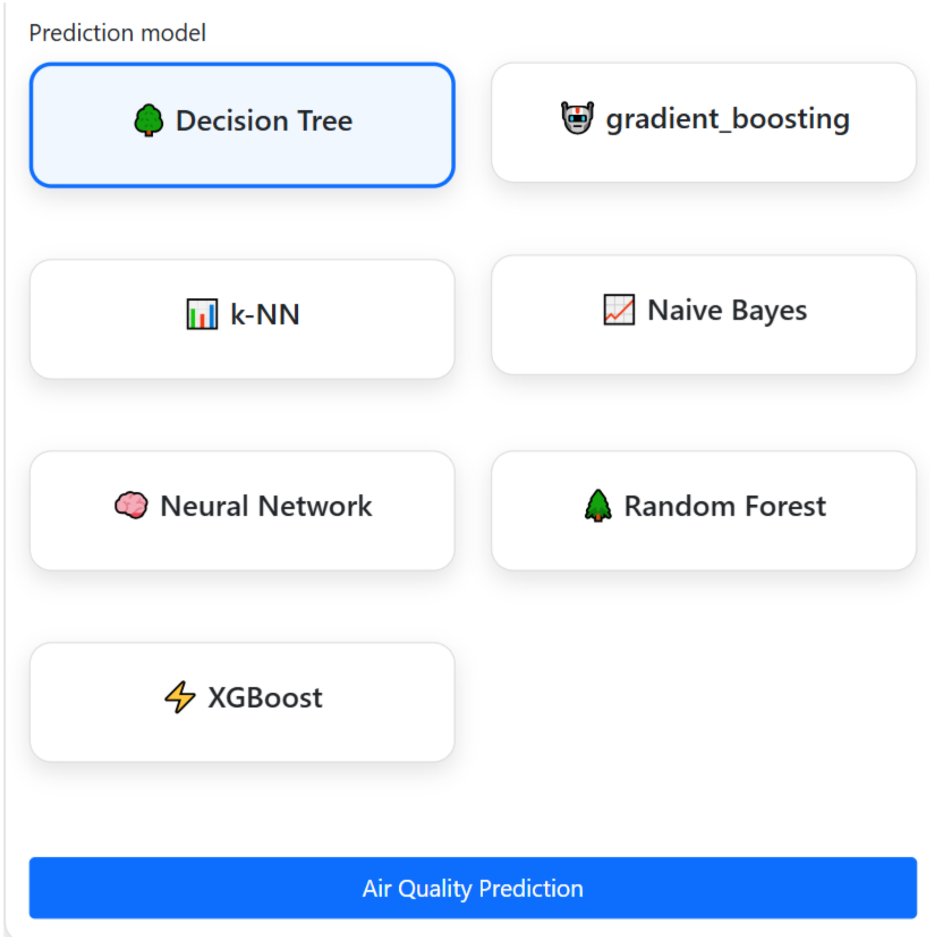


Figure 26: Overview of Prediction Models

3.9.1 Decision Tree

A Decision Tree is a type of supervised learning algorithm used for classification and regression tasks. It models decisions in a tree structure:

- **Root Node:** Represents the entire dataset.
- **Internal Nodes:** Represent tests on attributes.
- **Branches:** Outcomes of the tests.
- **Leaf Nodes:** Represent the final decision.

```python
'decision_tree': {
    'class': DecisionTreeClassifier,
    'params': {
        'max_depth': None,
        'min_samples_split': 2,
        'class_weight': 'balanced',
        'random_state': 42
    },
    'search_params': {
        'max_depth': [5, 10, None],
        'min_samples_split': [2, 5, 10],
        'criterion': ['gini', 'entropy']
    }
},
```

Figure 27: Code Example for Decision Tree Model

This interface provides the air quality prediction result using the Decision Tree model for the city of Delhi in India; the result is 'Moderate' on the date 2025/05/24.

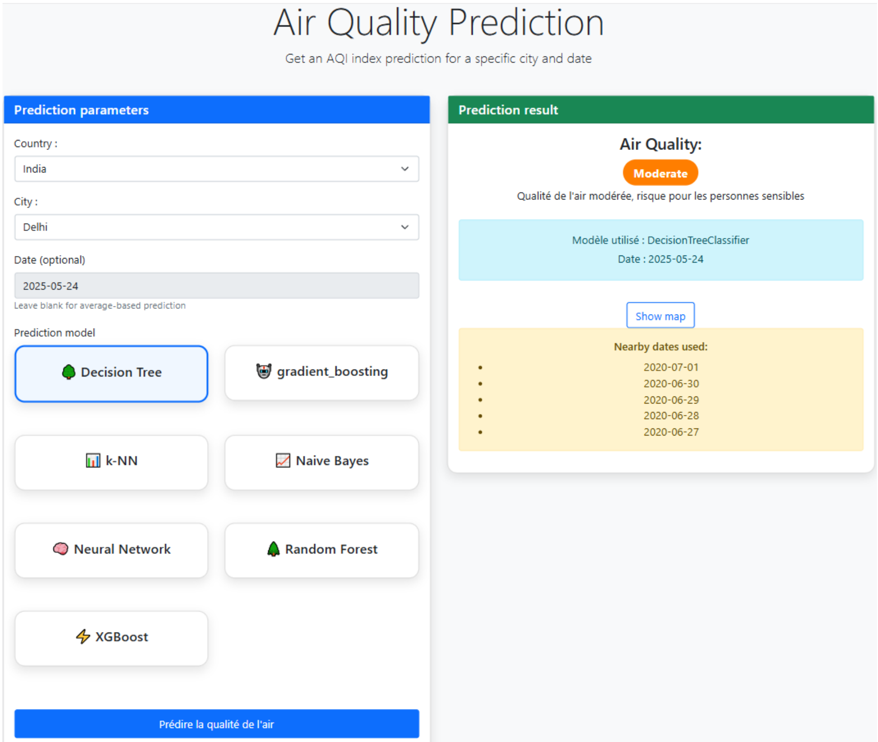


Figure 28: Prediction Result using Decision Tree Model

3.9.2 Gradient Boosting

Gradient Boosting builds models sequentially to correct previous errors:

1. Train a weak learner.
2. Compute residuals.
3. Fit new model to residuals.
4. Update ensemble.

```
'gradient_boosting': {
    'class': GradientBoostingClassifier,
    'params': {
        'n_estimators': 100,
        'learning_rate': 0.1,
        'max_depth': 3,
        'random_state': 42
    },
    'search_params': {
        'n_estimators': [50, 100],
        'learning_rate': [0.05, 0.1],
        'max_depth': [3, 5]
    }
}
```

Figure 29: Code Example for Gradient Boosting Model

This interface provides the air quality prediction result using the Gradient Boosting model for the city of Mumbai in India; the result is 'Satisfactory' on the date 2025/05/23.

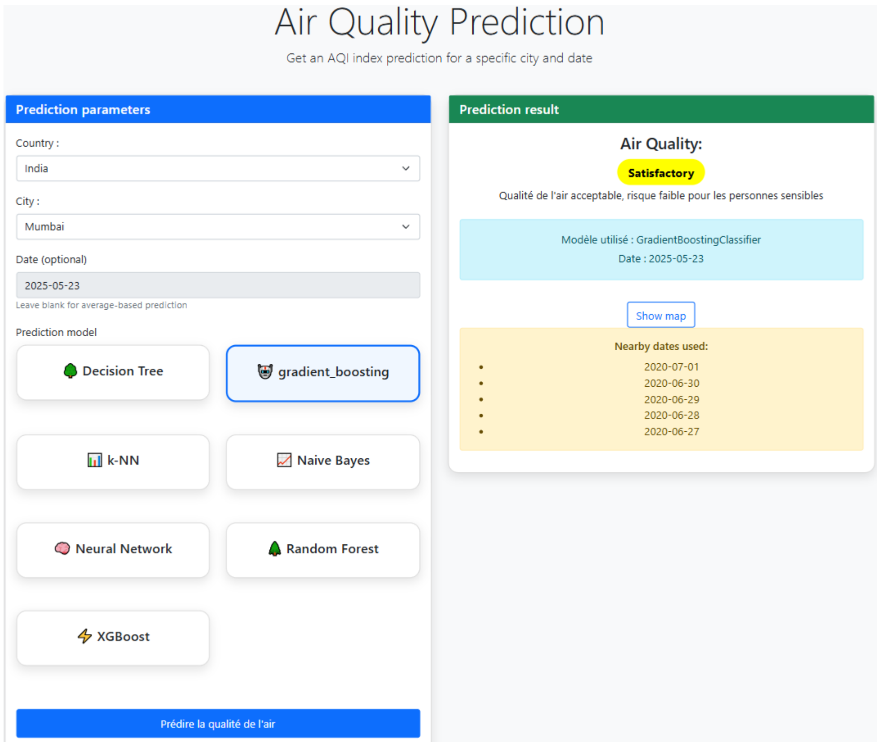


Figure 30: Prediction Result using Gradient Boosting Model

3.9.3 K-Nearest Neighbors (KNN)

KNN predicts based on the majority label of its k nearest neighbors:

1. Select k neighbors.
2. Compute distances.
3. Identify nearest k points.
4. Assign majority class.

```
'knn': {
    'class': KNeighborsClassifier,
    'params': {
        'n_neighbors': 5,
        'weights': 'uniform',
        'algorithm': 'auto',
        'n_jobs': -1
    },
    'search_params': {
        'n_neighbors': [3, 5, 7],
        'weights': ['uniform', 'distance'],
        'p': [1, 2]
    }
},
```

Figure 31: Code Example for KNN Model

This interface provides the air quality prediction result using the K-Nearest Neighbors algorithm for the city of Delhi in India; the result is 'Moderate' on the date 2025/05/22.

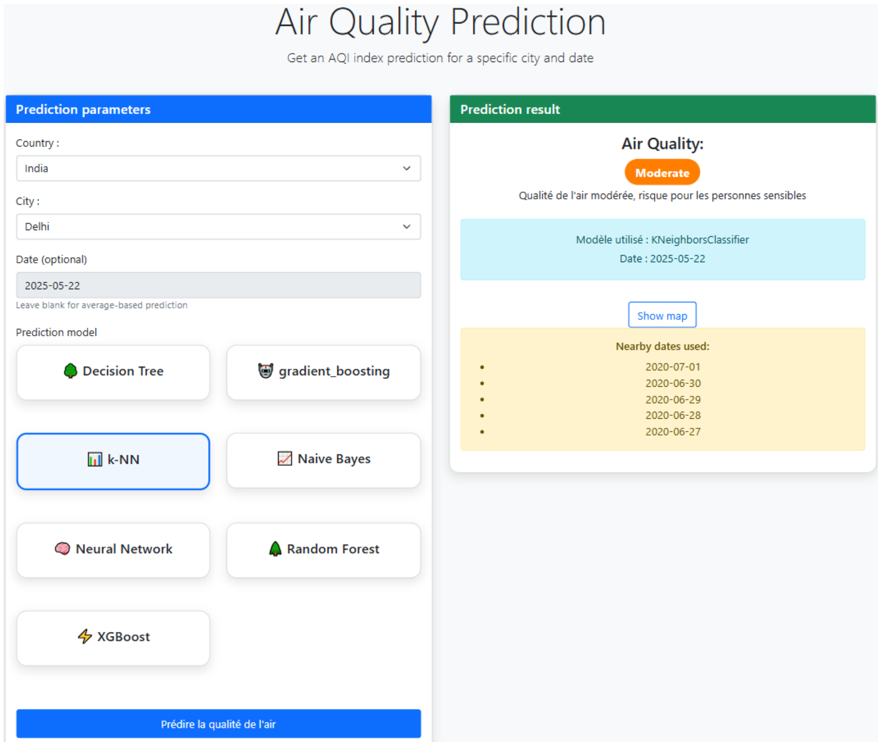


Figure 32: Prediction Result using KNN Model

3.9.4 Naive Bayes

Naive Bayes uses Bayes' theorem assuming feature independence:

- Computes class probabilities.
- Predicts the most probable class.

```
'naive_bayes': {
    'class': GaussianNB,
    'params': {},
    'search_params': {
        'var_smoothing': [1e-9, 1e-8, 1e-7]
    }
},
```

Figure 33: Code Example for Naive Bayes Model

This interface provides the air quality prediction result using the Naive Bayes algorithm for the city of Khenchela in Algeria; the result is 'Unhealthy' on the date 2025/05/21.

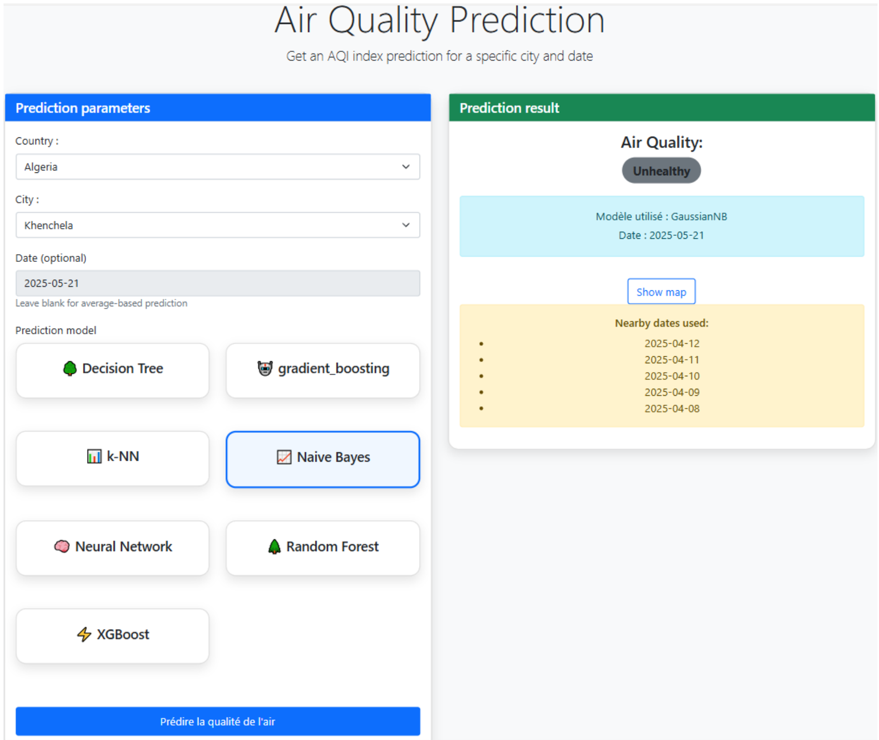


Figure 34: Prediction Result using Naive Bayes Model

3.9.5 Neural Network

Neural Networks simulate brain-like structures using layers of neurons:

1. Input layer processes features.
2. Hidden layers apply weighted transformations.
3. Output layer generates prediction.

```
'neural_network': {
    'class': MLPClassifier,
    'params': {
        'hidden_layer_sizes': (100,),
        'activation': 'relu',
        'solver': 'adam',
        'alpha': 0.0001,
        'batch_size': 'auto',
        'learning_rate': 'constant',
        'max_iter': 500,
        'random_state': 42,
        'early_stopping': True,
        'tol': 1e-4
    },
    'search_params': {
        'hidden_layer_sizes': [(50,), (100,), (50, 50)],
        'activation': ['relu', 'tanh'],
        'alpha': [0.0001, 0.001]
    }
},
```

Figure 35: Code Example for Neural Network Model

This interface provides the air quality prediction result using the Neural Network model for the city of Batna in Algeria; the result is 'Moderate' on the date 2025/05/21.

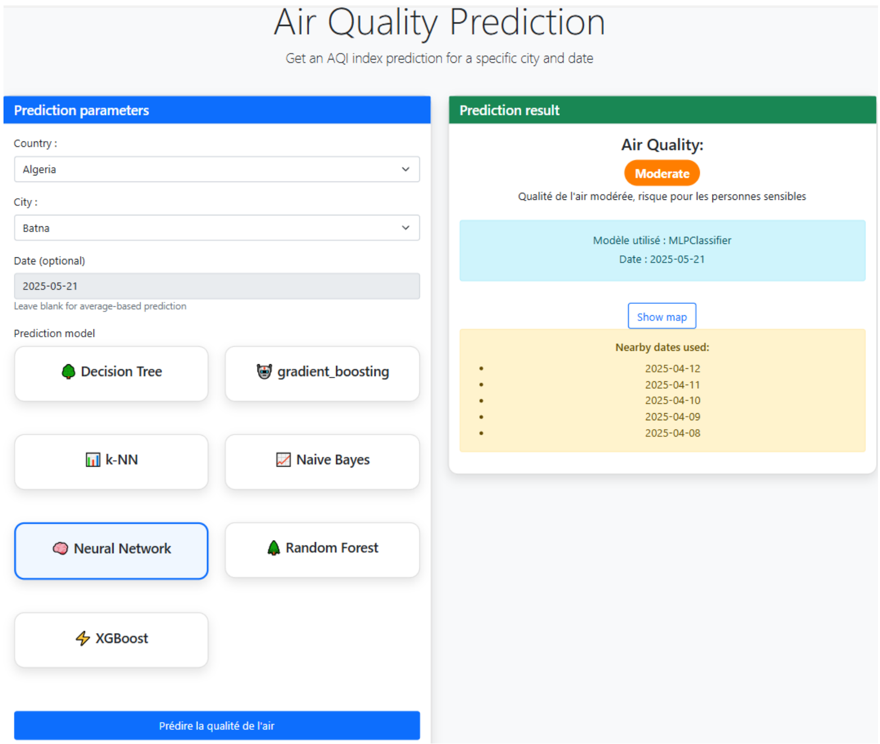


Figure 36: Prediction Result using Neural Network Model

3.9.6 Random Forest

Random Forest builds multiple decision trees and averages their results:

1. Train many trees on random subsets.
2. Aggregate predictions via voting or averaging.

```python
    'random_forest': {
        'class': RandomForestClassifier,
        'params': {
            'n_estimators': 200,
            'max_depth': 15,
            'min_samples_split': 5,
            'class_weight': 'balanced',
            'random_state': 42,
            'n_jobs': -1
        },
        'search_params': {
            'n_estimators': [100, 200],
            'max_depth': [10, 15],
            'min_samples_split': [5, 10]
        }
    },
```

Figure 37: Code Example for Random Forest Model

This interface provides the air quality prediction result using the Random Forest model for the city of Constantine in Algeria; the result is 'Satisfactory' on the date 2025/04/10.

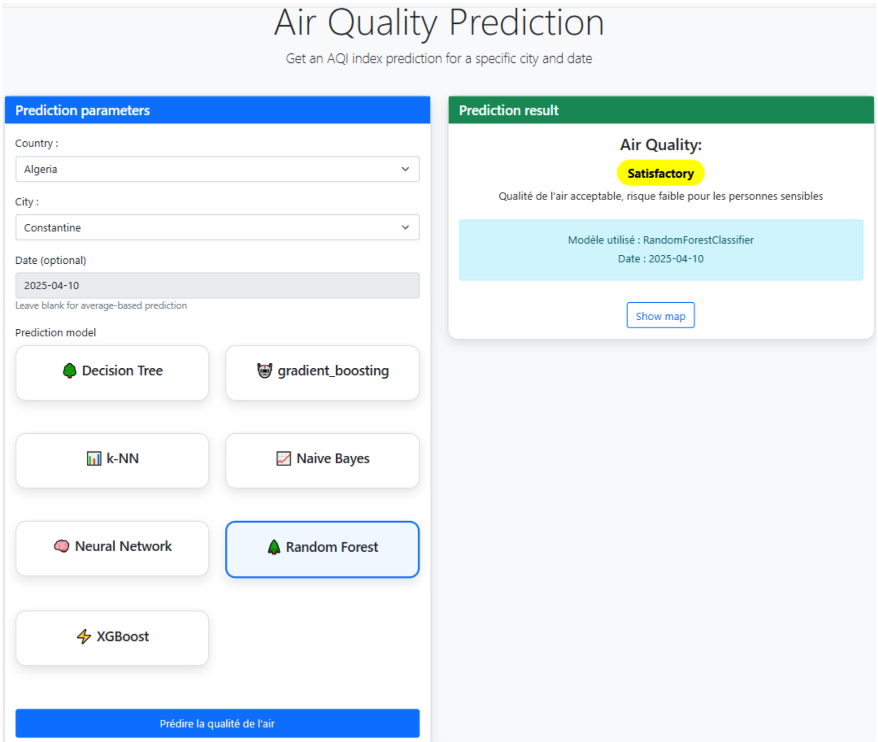


Figure 38: Prediction Result using Random Forest Model

3.9.7 XGBoost

XGBoost uses optimized gradient boosting techniques:

1. Sequentially adds trees correcting prior errors.
2. Applies regularization to prevent overfitting.

```python
'xgboost': {
    'class': XGBClassifier,
    'params': {
        'n_estimators': 150,
        'max_depth': 6,
        'learning_rate': 0.1,
        'subsample': 0.8,
        'colsample_bytree': 0.8,
        'objective': 'multi:softmax',
        'random_state': 42,
        'use_label_encoder': False,
        'eval_metric': 'mlogloss',
        'n_jobs': -1
    },
    'search_params': {
        'n_estimators': [100, 150],
        'max_depth': [6, 8],
        'learning_rate': [0.1, 0.2]
    }
},
```

Figure 39: Code Example for XGBoost Model

This interface provides the air quality prediction result using the XGBoost model for the city of Ouargla in Algeria; the result is 'Satisfactory' on the date 2025/04/10.

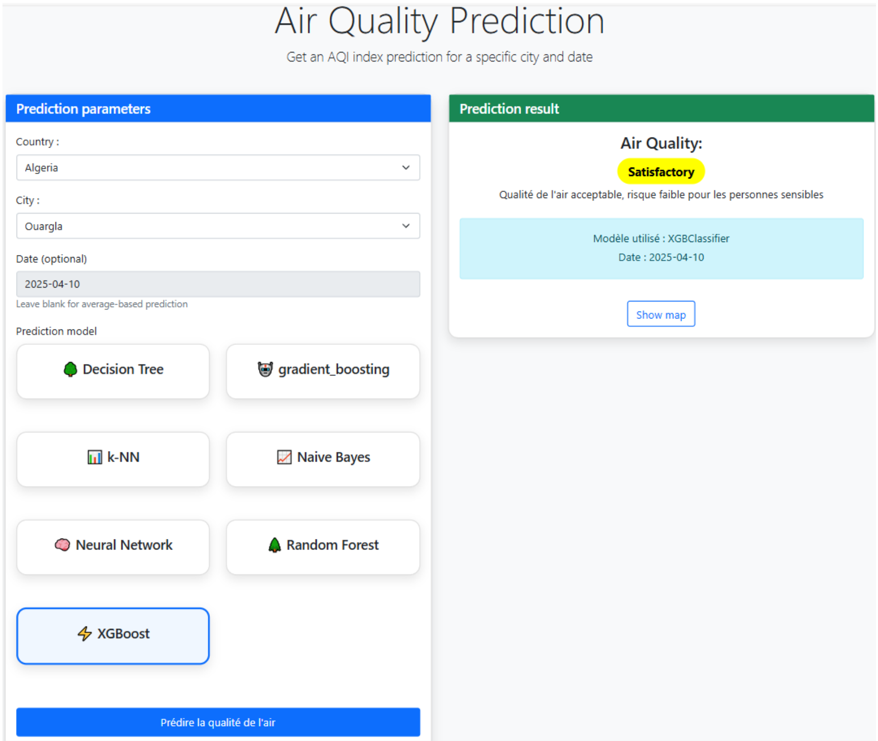


Figure 40: Prediction Result using XGBoost Model

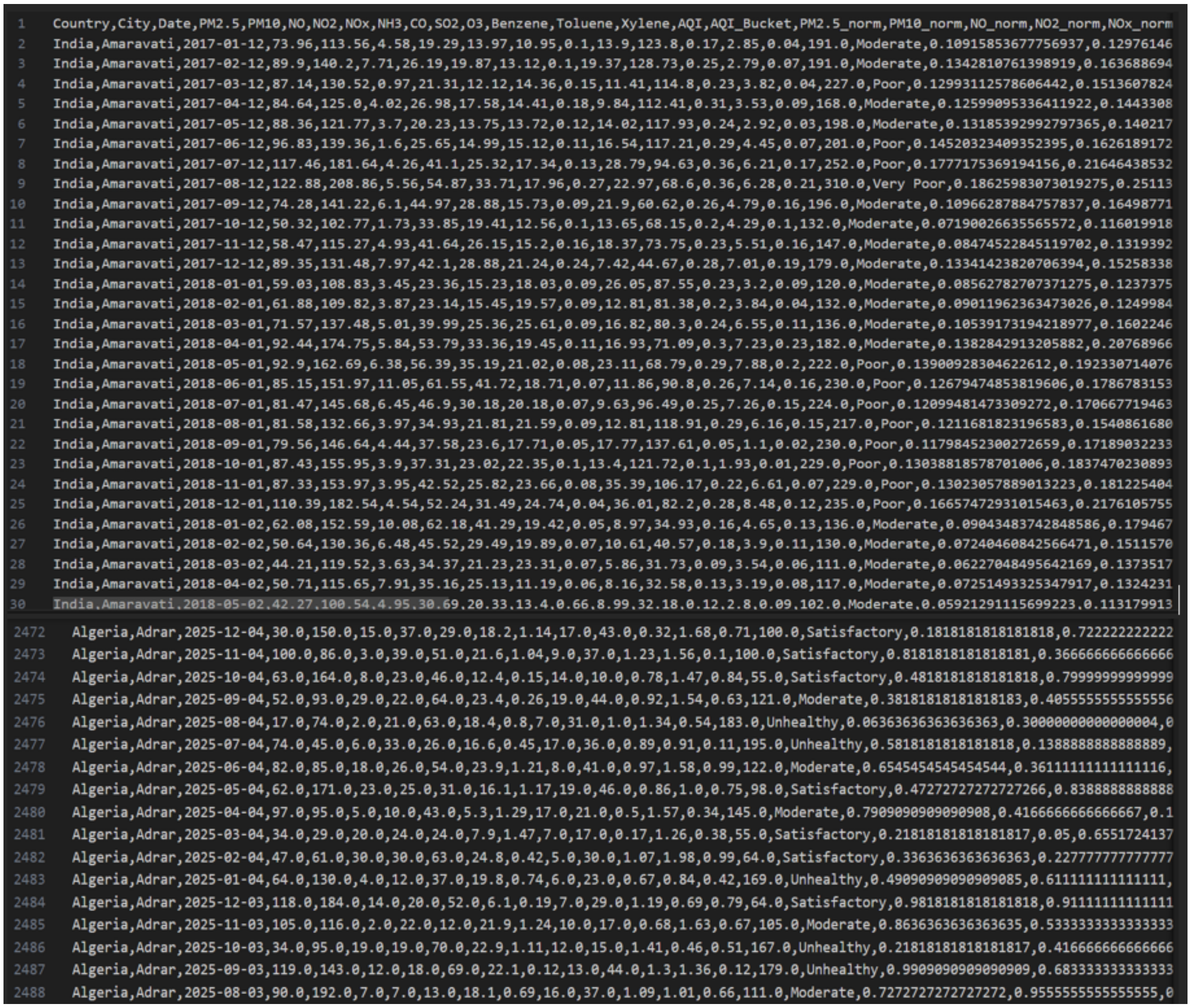
3.10 Dataset Description and Utilization

The dataset employed for this project is a combination of historical air quality measurements from multiple global locations, including cities in India and Algeria. The dataset contains a wide range of environmental and pollutant-related features that are essential for predicting air quality levels.

Dataset Features: Each record in the dataset includes the following variables:

- **Country, City, Date:** Geographic and temporal information.
- **PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene:** Concentrations of various air pollutants.
- **AQI (Air Quality Index):** A numerical index representing overall air quality.
- **AQI Bucket:** Categorical air quality level (e.g., Moderate, Poor, Satisfactory).
- **Normalized Columns:** Features such as PM2.5_norm, PM10_norm, NO_norm, NO2_norm, etc., are scaled to ensure consistent data range for model training.

Data Sample: A snippet of the dataset is illustrated in Figure 41.



```
Country,City,Date,PM2.5,PM10,NO,NO2,NOx,NH3,CO,SO2,O3,Benzene,Toluene,Xylene,AQI,AQI_Bucket,PM2.5_norm,PM10_norm,NO_norm,NO2_norm,NOx_norm
India,Amaravati,2017-01-12,73.96,113.56,4.58,19.29,13.97,10.95,0.1,13.9,123.8,0.17,2.85,0.04,191.0,Moderate,0.10915853677756937,0.12976146
India,Amaravati,2017-02-12,89.9,140.2,7.71,26.19,19.87,13.12,0.1,19.37,128.73,0.25,2.79,0.07,191.0,Moderate,0.1342810761398919,0.163688694
India,Amaravati,2017-03-12,87.14,130.52,0.97,21.31,12.12,14.36,0.15,11.41,114.8,0.23,3.82,0.04,227.0,Poor,0.12993112578606442,0.1513607824
India,Amaravati,2017-04-12,84.64,125.0,4.02,26.98,17.58,14.41,0.18,9.84,112.41,0.31,3.53,0.09,168.0,Moderate,0.12599095336411922,0.1443308
India,Amaravati,2017-05-12,88.36,121.77,3.7,20.23,13.75,13.72,0.12,14.02,117.93,0.24,2.92,0.03,198.0,Moderate,0.13185392992797365,0.140217
India,Amaravati,2017-06-12,96.83,139.36,1.6,25.65,14.99,15.12,0.11,16.54,117.21,0.29,4.45,0.07,201.0,Poor,0.1452032340935239,0.1626189172
India,Amaravati,2017-07-12,117.46,181.64,4.26,41.1,25.32,17.34,0.13,28.79,94.63,0.36,6.21,0.17,252.0,Poor,0.177717536919415,0.21646438532
India,Amaravati,2017-08-12,122.88,208.86,5.56,54.87,33.71,17.96,0.27,22.97,68.6,0.36,6.28,0.21,310.0,Very Poor,0.18625983073019275,0.25113
India,Amaravati,2017-09-12,74.28,141.22,6.1,44.97,28.88,15.73,0.09,21.9,60.62,0.26,4.79,0.16,196.0,Moderate,0.10966287884757837,0.16498771
India,Amaravati,2017-10-12,50.32,102.77,1.73,33.85,19.41,12.56,0.1,13.65,68.15,0.2,4.29,0.1,132.0,Moderate,0.07190026635565572,0.116019918
India,Amaravati,2017-11-12,58.47,115.27,4.93,41.64,26.15,15.2,0.16,18.37,73.75,0.23,5.51,0.16,147.0,Moderate,0.08474522845119702,0.1319392
India,Amaravati,2017-12-12,89.35,131.48,7.97,42.1,28.88,21.24,0.24,7.42,44.67,0.28,7.01,0.19,179.0,Moderate,0.13341423820706394,0.15258338
India,Amaravati,2018-01-01,59.03,108.83,3.45,23.36,15.23,18.03,0.09,26.05,87.55,0.23,3.2,0.09,120.0,Moderate,0.08562782707371275,0.1237375
India,Amaravati,2018-02-01,61.88,109.82,3.87,23.14,15.45,19.57,0.09,12.81,81.38,0.2,3.84,0.04,132.0,Moderate,0.09011962363473026,0.1249984
India,Amaravati,2018-03-01,71.57,137.48,5.01,39.99,25.36,25.61,0.09,16.82,80.3,0.24,6.55,0.11,136.0,Moderate,0.1053917319421897,0.1602246
India,Amaravati,2018-04-01,92.44,174.75,5.84,53.79,33.36,19.45,0.11,16.93,71.09,0.3,7.23,0.23,182.0,Moderate,0.138284291320588,0.20768966
India,Amaravati,2018-05-01,92.9,162.69,6.38,56.39,35.19,21.02,0.08,23.11,68.79,0.29,7.88,0.2,222.0,Poor,0.1390092830462261,0.192330714076
India,Amaravati,2018-06-01,85.15,151.97,11.05,61.55,41.72,18.71,0.07,11.86,90.8,0.26,7.14,0.16,230.0,Poor,0.1267947485381960,0.1786783153
India,Amaravati,2018-07-01,81.47,145.68,6.45,46.9,30.18,20.18,0.07,9.63,96.49,0.25,7.26,0.15,224.0,Poor,0.1209948147330927,0.170667719463
India,Amaravati,2018-08-01,81.58,132.66,3.97,34.93,21.81,21.59,0.09,12.81,118.91,0.29,6.16,0.15,217.0,Poor,0.121168182319658,0.1540861680
India,Amaravati,2018-09-01,79.56,146.64,4.44,37.58,23.6,17.71,0.05,17.77,137.61,0.05,1.1,0.02,230.0,Poor,0.1179845230027265,0.17189032233
India,Amaravati,2018-10-01,87.43,155.95,3.9,37.31,23.02,22.35,0.1,13.4,121.72,0.1,1.93,0.01,229.0,Poor,0.13038818578701006,0.1837470230893
India,Amaravati,2018-11-01,87.33,153.97,3.95,42.52,25.82,23.66,0.08,35.39,106.17,0.22,6.61,0.07,229.0,Poor,0.1302305788901322,0.181225404
India,Amaravati,2018-12-01,110.39,182.54,4.54,52.24,31.49,24.74,0.04,36.01,82.2,0.28,8.48,0.12,235.0,Poor,0.1665747293101546,0.2176105755
India,Amaravati,2018-01-02,62.08,152.59,10.08,62.18,41.29,19.42,0.05,8.97,34.93,0.16,4.65,0.13,136.0,Moderate,0.09043483742848586,0.179467
India,Amaravati,2018-02-02,50.64,130.36,6.48,45.52,29.49,19.89,0.07,10.61,40.57,0.18,3.9,0.11,130.0,Moderate,0.0724046084256647,0.1511570
India,Amaravati,2018-03-02,44.21,119.52,3.63,34.37,21.23,23.31,0.07,5.86,31.73,0.09,3.54,0.06,111.0,Moderate,0.06227048495642169,0.1373517
India,Amaravati,2018-04-02,50.71,115.65,7.91,35.16,25.13,11.19,0.06,8.16,32.58,0.13,3.19,0.08,117.0,Moderate,0.07251493325347917,0.1324231
India,Amaravati,2018-05-02,42.27,100.54,4.95,30.69,20.33,13.4,0.66,8.99,32.18,0.12,2.8,0.09,102.0,Moderate,0.0592129111569922,0.113179913
Algeria,Adrar,2025-12-04,30.0,150.0,15.0,37.0,29.0,18.2,1.14,17.0,43.0,0.32,1.68,0.71,100.0,Satisfactory,0.1818181818181818,0.72222222222
Algeria,Adrar,2025-11-04,100.0,86.0,3.0,39.0,51.0,21.6,1.04,9.0,37.0,1.23,1.56,0.1,100.0,Satisfactory,0.818181818181818,0.36666666666666
Algeria,Adrar,2025-10-04,63.0,164.0,8.0,23.0,46.0,12.4,0.15,14.0,10.0,0.78,1.47,0.84,55.0,Satisfactory,0.4818181818181818,0.7999999999999
Algeria,Adrar,2025-09-04,52.0,93.0,29.0,22.0,64.0,23.4,0.26,19.0,44.0,0.92,1.54,0.63,121.0,Moderate,0.38181818181818183,0.4055555555555556
Algeria,Adrar,2025-08-04,17.0,74.0,2.0,21.0,63.0,18.4,0.8,7.0,31.0,1.0,1.34,0.54,183.0,Unhealthy,0.0636363636363636,0.30000000000000004,0
Algeria,Adrar,2025-07-04,74.0,45.0,6.0,33.0,26.0,16.6,0.45,17.0,36.0,0.89,0.91,0.11,195.0,Unhealthy,0.5818181818181818,0.1388888888888889,
Algeria,Adrar,2025-06-04,82.0,85.0,18.0,26.0,54.0,23.9,1.21,8.0,41.0,0.97,1.58,0.99,122.0,Moderate,0.654545454545454,0.3611111111111116,
Algeria,Adrar,2025-05-04,62.0,171.0,23.0,25.0,31.0,16.1,1.17,19.0,46.0,0.86,1.0,0.75,98.0,Satisfactory,0.4727272727272727266,0.8388888888888
Algeria,Adrar,2025-04-04,97.0,95.0,5.0,10.0,43.0,5.3,1.29,17.0,21.0,0.5,1.57,0.34,145.0,Moderate,0.790909090909090908,0.4166666666666667,0.1
Algeria,Adrar,2025-03-04,34.0,29.0,20.0,24.0,24.0,7.9,1.47,7.0,17.0,0.17,1.26,0.38,55.0,Satisfactory,0.21818181818181817,0.05,0.6551724137
Algeria,Adrar,2025-02-04,47.0,61.0,30.0,30.0,63.0,24.8,0.42,5.0,30.0,1.07,1.98,0.99,64.0,Satisfactory,0.336363636363636363,0.22777777777777
Algeria,Adrar,2025-01-04,64.0,130.0,4.0,12.0,37.0,19.8,0.74,6.0,23.0,0.67,0.84,0.42,169.0,Unhealthy,0.4909090909090909085,0.611111111111111,
Algeria,Adrar,2025-12-03,118.0,184.0,14.0,20.0,52.0,6.1,0.19,7.0,29.0,1.19,0.69,0.79,64.0,Satisfactory,0.981818181818181818,0.91111111111111
Algeria,Adrar,2025-11-03,105.0,116.0,2.0,22.0,12.0,21.9,1.24,10.0,17.0,0.68,1.63,0.67,105.0,Moderate,0.863636363636363635,0.5333333333333333
Algeria,Adrar,2025-10-03,34.0,95.0,19.0,19.0,70.0,22.9,1.11,12.0,15.0,1.41,0.46,0.51,167.0,Unhealthy,0.21818181818181817,0.41666666666666
Algeria,Adrar,2025-09-03,119.0,143.0,12.0,18.0,69.0,22.1,0.12,13.0,44.0,1.3,1.36,0.12,179.0,Unhealthy,0.9909090909090909,0.68333333333333
Algeria,Adrar,2025-08-03,90.0,192.0,7.0,7.0,13.0,18.1,0.69,16.0,37.0,1.09,1.01,0.66,111.0,Moderate,0.727272727272727272,0.955555555555555,0
```

Figure 41: Sample records from the air quality dataset.

Data Preprocessing Steps:

1. **Missing Value Handling:** Missing data was imputed using mean substitution for continuous variables.
2. **Normalization:** Features were scaled between 0 and 1 using Min-Max normalization.
3. **Categorical Encoding:** AQI Buckets were label encoded for use in classification models.
4. **Splitting:** The dataset was divided into 70% training data and 30% testing data.

Dataset Utilization: The processed dataset served as input to all machine learning models discussed in Section 3.9. Each model learned from these features to predict the AQI Bucket for unseen data points. The variety in pollutant measures ensured robust training and better generalization capability of the models.

This dataset played a crucial role in validating the performance and reliability of the air quality prediction system developed in this work.

3.11 Working Demonstration

Our developed system offers three main functionalities that demonstrate its versatility and practical applicability:

- **City Prediction:** Predict air quality for a selected city based on real-time and historical data.
- **Manual Prediction:** Allow the user to input environmental parameters manually for prediction.
- **AQI Detection via API:** Retrieve and predict air quality using live API data sources.

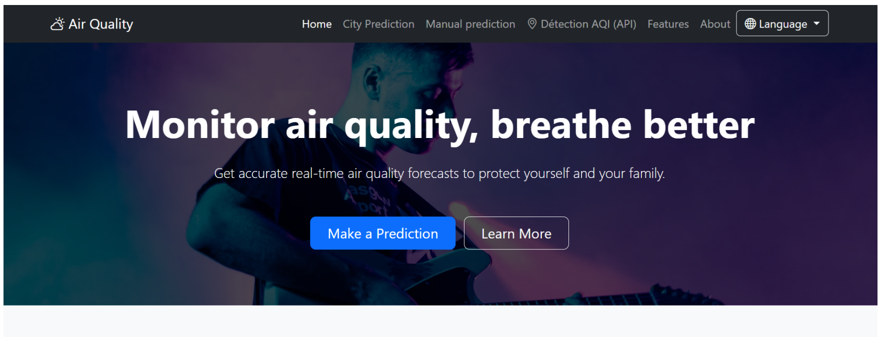


Figure 42: Home page of the application.

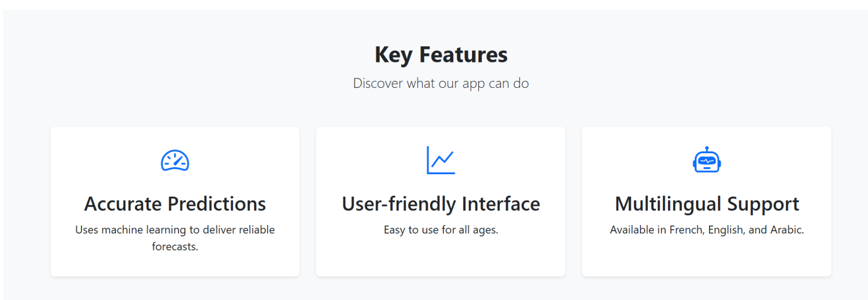


Figure 43: Discover what the app can do.

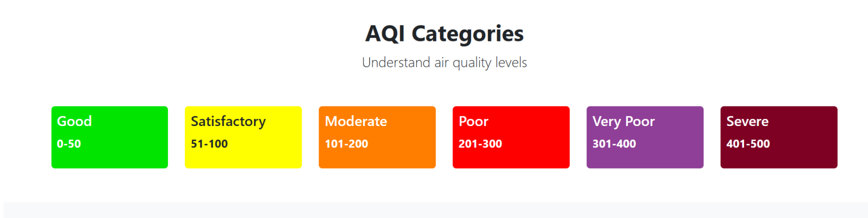


Figure 44: Air quality levels displayed in the application.

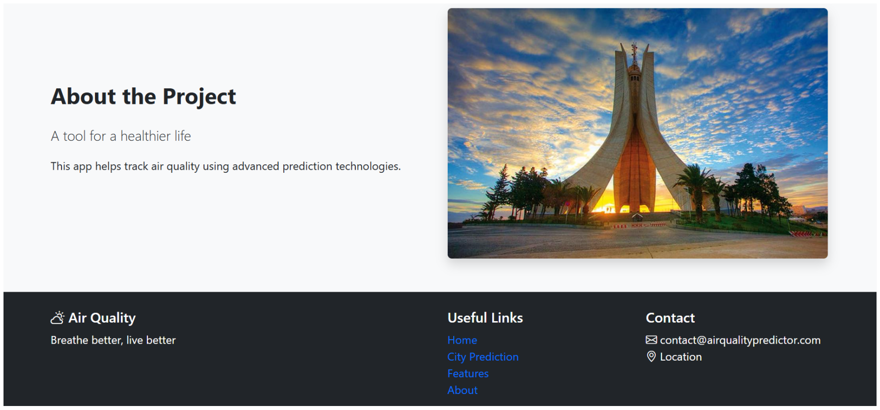


Figure 45: About the Project section.

- `City Prediction:` `Predict air quality for a selected city based on real-time and historical data.`

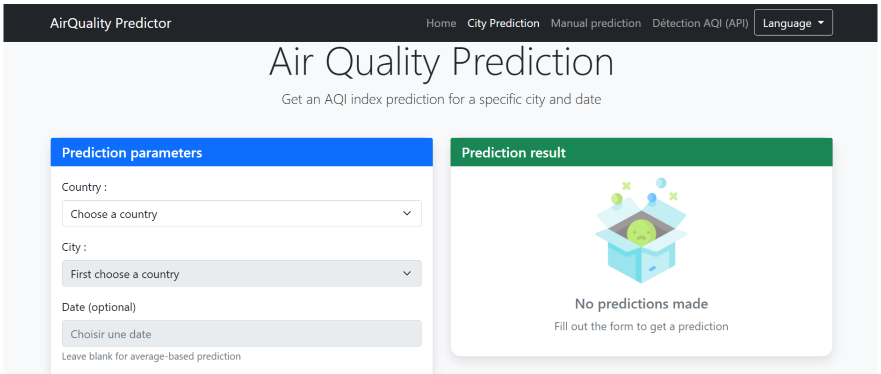


Figure 46: Air Quality Prediction interface.

This interface provides the air quality prediction result using the Random Forest model for the city of Mumbai in India; the result is 'Satisfactory' on the date 2025/05/10.

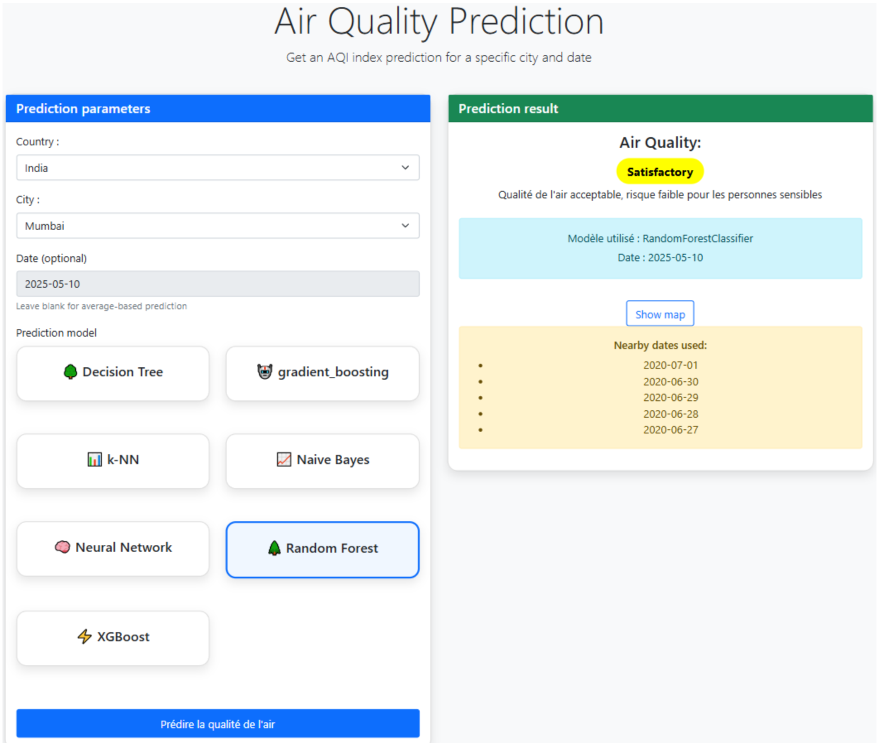


Figure 47: Prediction result for the city of Mumbai, India.

This image shows the location of Mumbai city when we click on the "Show Map" button in Figure 47.

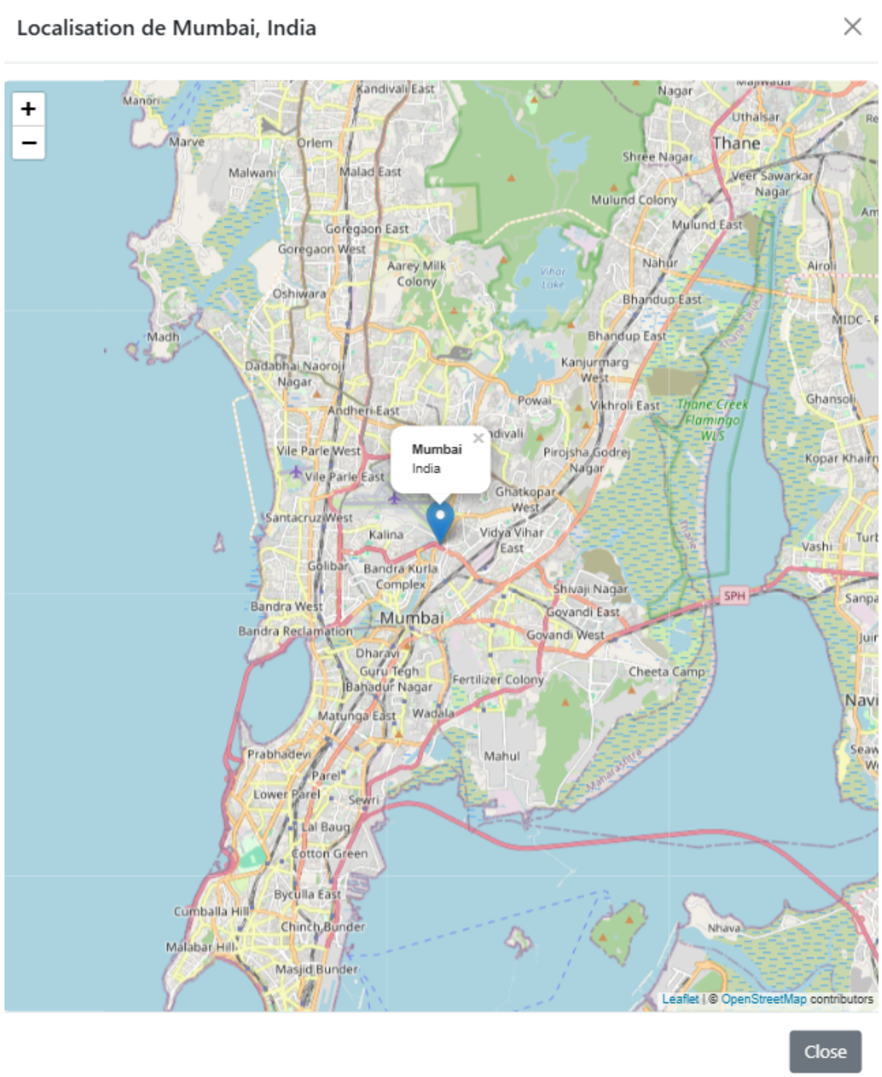


Figure 48: Map location for the city of Mumbai, India.

This interface provides the air quality prediction result using the XGBoost algorithm for the city of Khenchela in Algeria; the result is 'Moderate' on the date 2025/05/10.

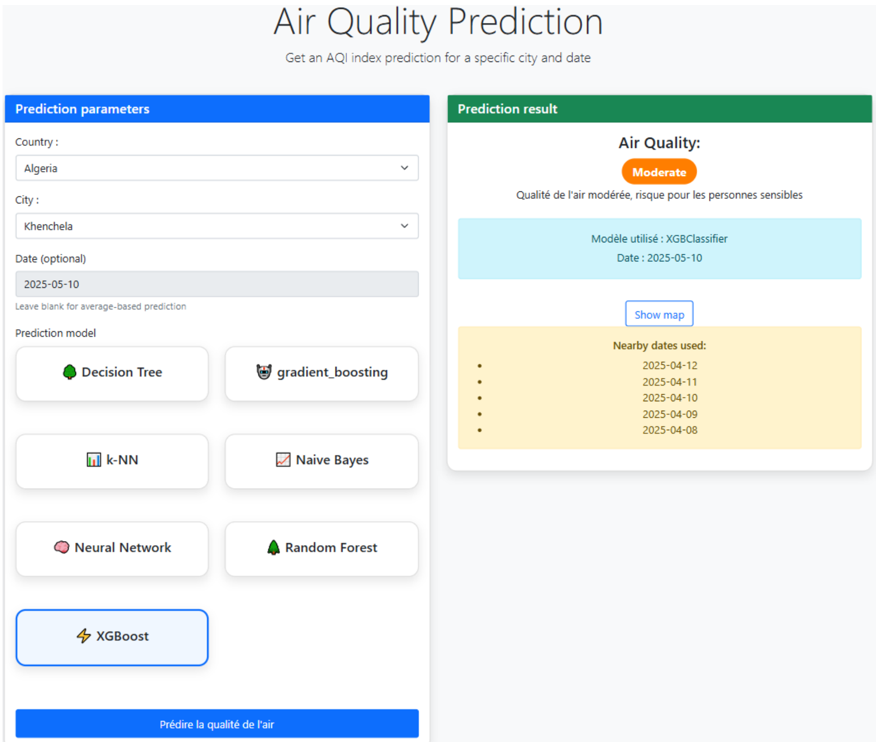


Figure 49: Prediction result for the city of Khenchela, Algeria.

This image shows the location of Khenchela city when we click on the "Show Map" button in Figure 49.

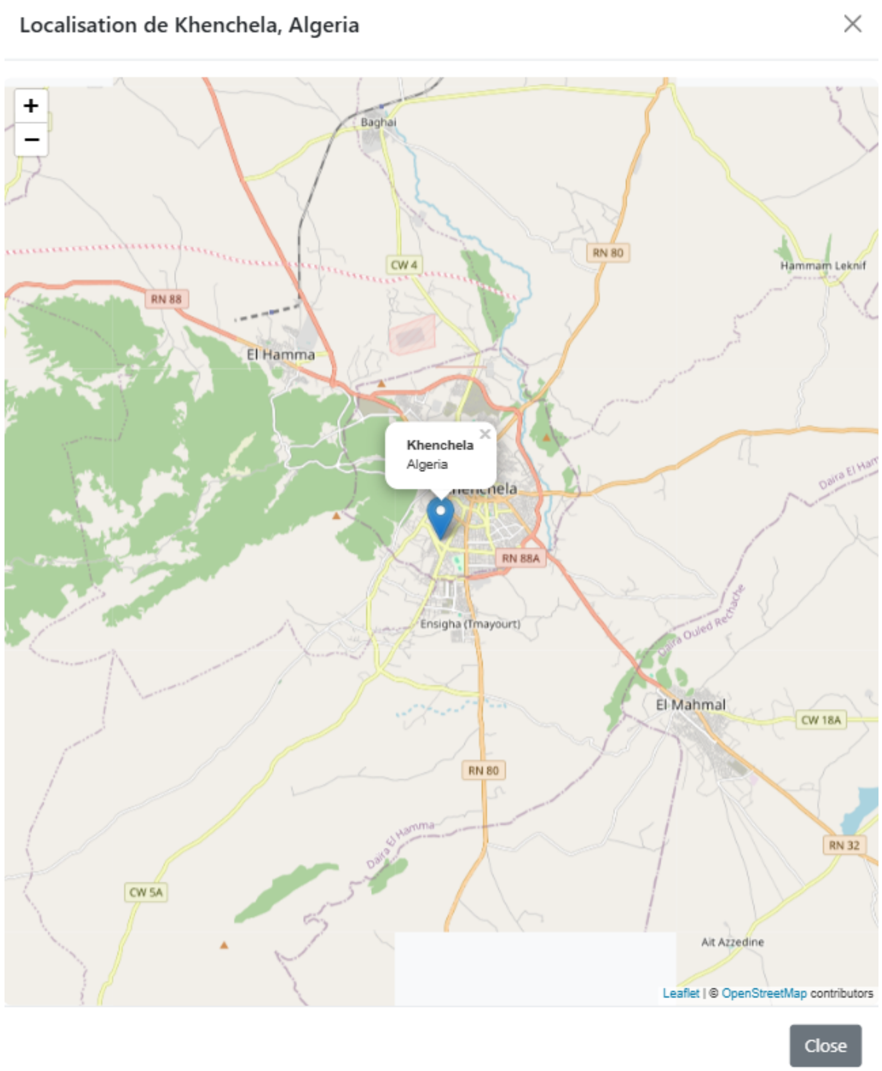


Figure 50: Map location for the city of Khenchela, Algeria.

- `Manual Prediction:` `Allow the user to input environmental parameters manually for prediction.`

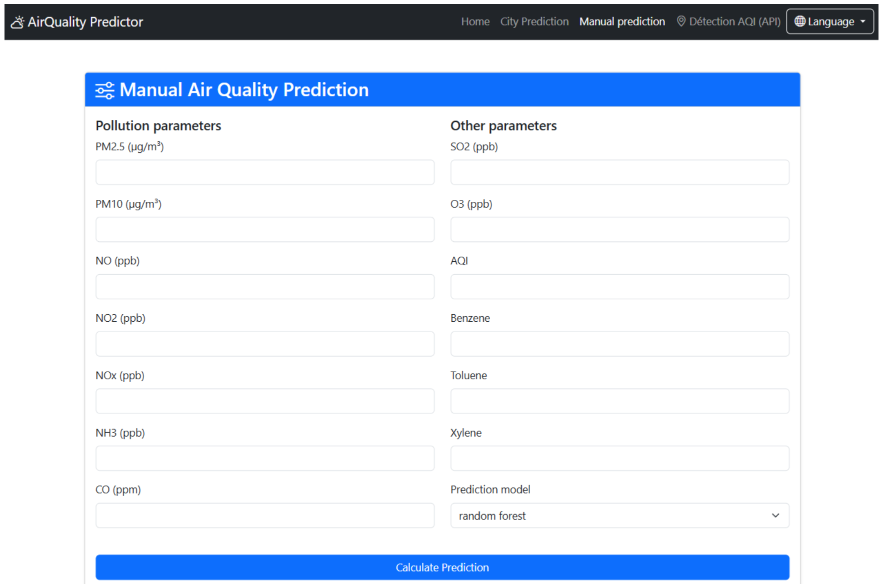


Figure 51: Manual Air Quality Prediction interface.

This interface provides the predicted air quality result manually by entering pollution parameters and other parameters using a Random Forest model; the result indicates good air quality.

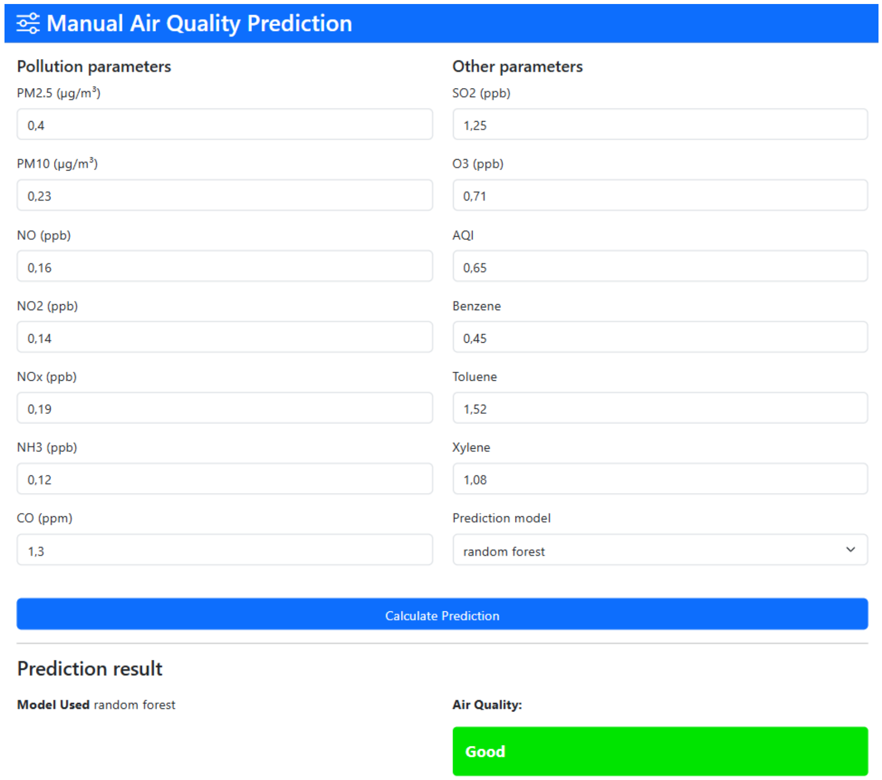


Figure 52: Result of Manual Air Quality Prediction.

- `AQI Detection via API`: Retrieve and predict air quality using live API data sources.

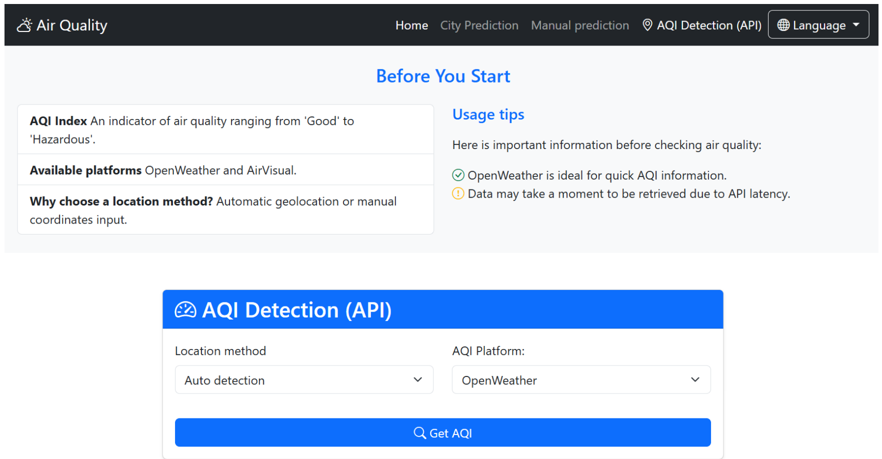


Figure 53: AQI Detection using external API sources.

This interface provides the predicted air quality result through automatic detection using an air quality monitoring platform AirVisual API; the result indicates moderate air quality. You can also click the "Show Details" button to view pollution parametrs and their value.

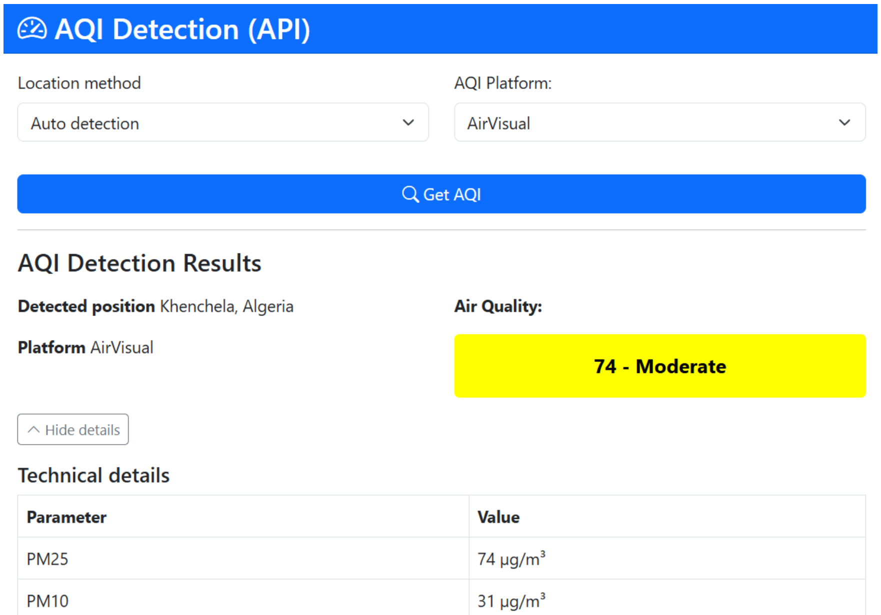


Figure 54: Auto detection result for Khenchela using AirVisual API.

This interface provides the predicted air quality result through manual detection using an air quality monitoring platform openWeather API; by entering the latitude and longitude of khenchela city. The result indicates good air quality, and you can also click the "Show Details" button to view the pollution parametrs and their values.

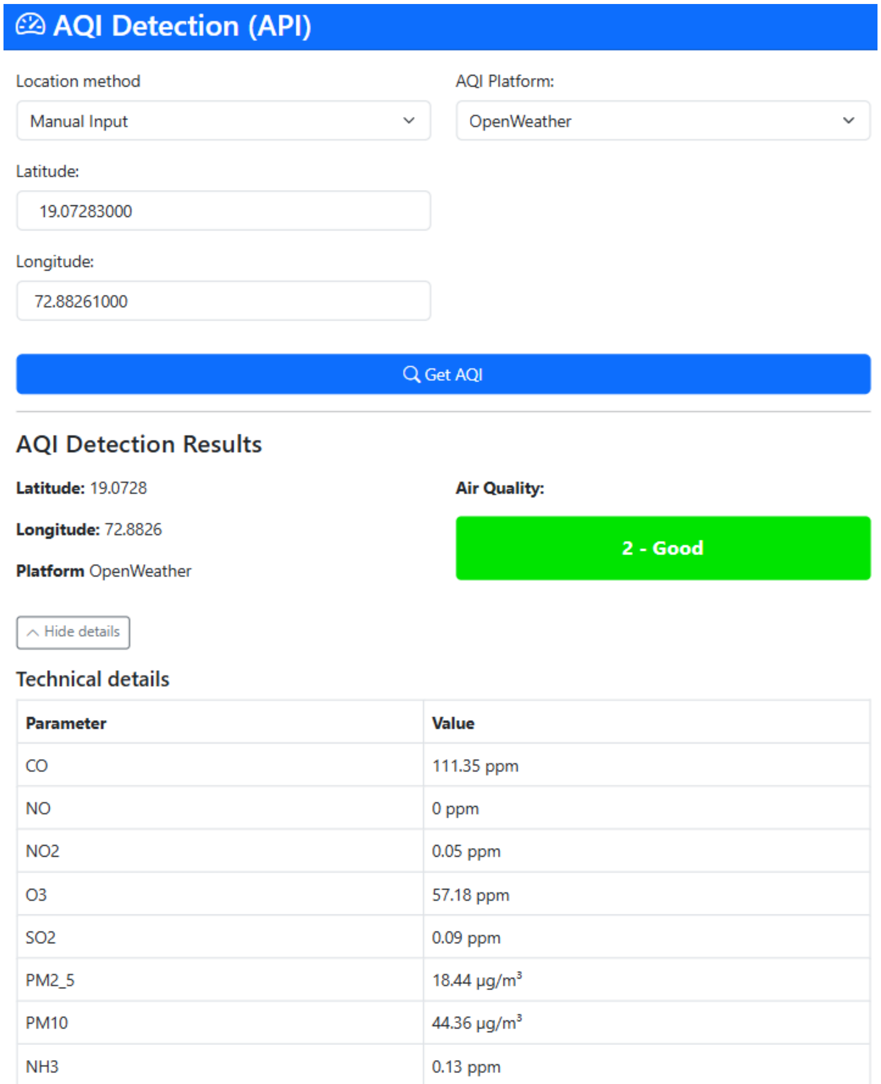


Figure 55: Manual input result for Khenchela using OpenWeather API.

This interface provides the predicted air quality result through manual detection using an air quality monitoring platform airvisual API; by entering the latitude and longitude of mombai city. The result indicates moderate air quality, and you can also click the "Show Details" button to view the pollution parametrs and their values.

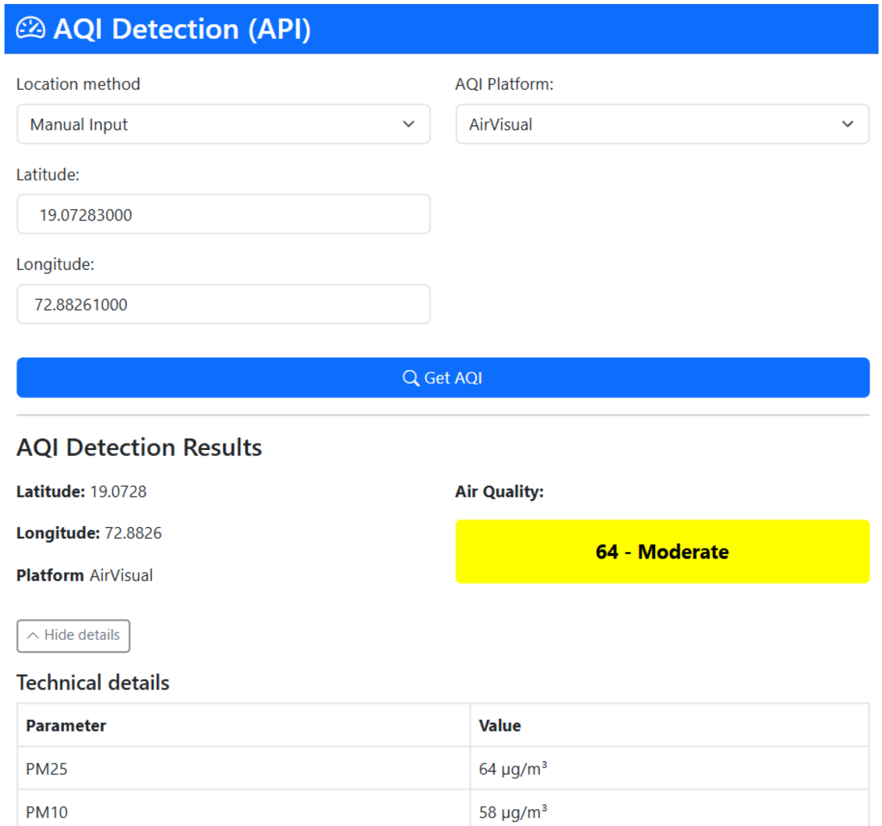


Figure 56: Manual input result for Mumbai using AirVisual API.

This interface provides the predicted air quality result through manual detection using an air quality monitoring platform openWeather API; by entering the latitude and longitude of mombai city. The result indicates good air quality, and you can also click the "Show Details" button to view the pollution parametrs and their values.

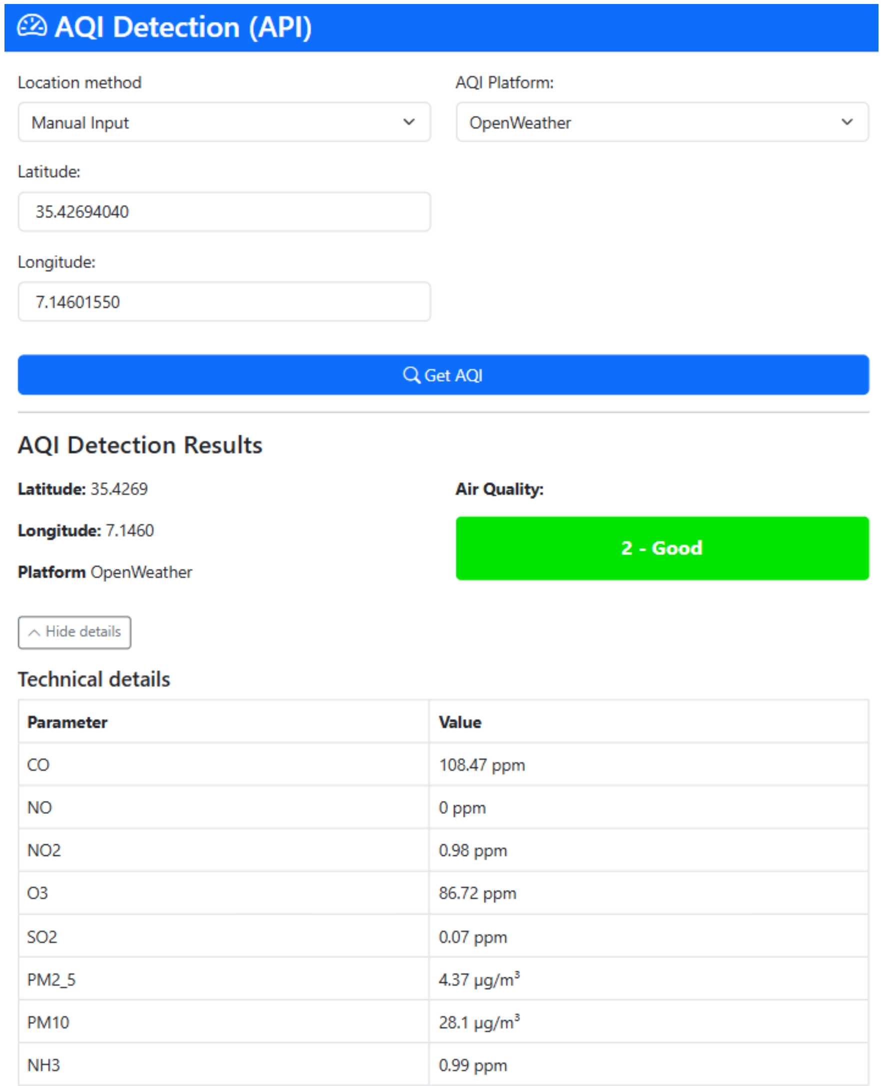


Figure 57: Manual input result for Mumbai using OpenWeather API.

3.12 How to Run the Application

To run the developed air quality prediction application, the following steps must be followed:

1. **Extract the Project Files:**
   Unzip the provided project archive to a directory of your choice.
2. **Create a Virtual Environment:**
   Execute the following command to create an isolated Python environment:

```
python -m venv venv
```

This ensures that the project dependencies are separated from the global Python environment.

3. **Allow Script Execution (For Windows PowerShell Users):**
   Temporarily allow script execution using:

   ```
   Set-ExecutionPolicy RemoteSigned -Scope Process
   ```

4. **Activate the Virtual Environment:**

   ```
   venv\Scripts\activate
   ```

   Once activated, Python and pip commands will operate within this environment.

5. **Upgrade pip:**

   ```
   python.exe -m pip install --upgrade pip
   ```

6. **Install Required Packages:**

   ```
   pip install -r requirements.txt
   ```

   This will install all libraries required for the application.

7. **Data Cleaning:**

   ```
   python src/data_processing.py
   ```

   This step processes and cleans raw data files.

8. **Exploratory Data Analysis (EDA):**

   ```
   python src/EDA.py
   ```

   This script generates visualizations and statistical summaries to understand the dataset better.

9. **Model Training:**

   ```
   python src/Model_Training.py
   ```

   This trains the machine learning models on the cleaned dataset.

10. **Optional: Test the Prediction System:**

    ```
    python src/predict.py
    ```

    Tests the trained models to validate their performance on sample inputs.

11. **Launch the Web Application:**

```
python src/app.py
```

Starts the web application locally.

12. **Access the Application:**
    Open your browser and navigate to:

    ```
    http://localhost:5000
    ```

    The web interface provides an interactive way to input data and view predictions.

Summary: This complete workflow includes environment setup, data cleaning, exploratory analysis, model training, testing, and deploying a functional machine learning web application.

4.Conclusion

This chapter comprehensively detailed the implementation of our air quality prediction system. We outlined the development environment, including hardware specifications and the selection of Visual Studio Code, Python, and relevant libraries. The graphical interfaces presented showcased the system's core functionalities, from predicting air quality in specific cities to demonstrating manual predictions and API integrations. While challenges were encountered, effective strategies ensured their resolution, resulting in a robust and functional application. This chapter, therefore, validates the successful practical realization of our air quality prediction system.

General Conclusion

Throughout this thesis, we have examined the complex and multifaceted issue of air pollution and explored the potential of artificial intelligence to mitigate its impacts. Beginning with a comprehensive scientific review of major air pollutants and their effects, we underscored the importance of accurate and timely air quality monitoring.

We then investigated the role of AI in environmental forecasting, evaluating a range of machine learning and deep learning techniques capable of predicting pollutant levels with high accuracy. Our practical implementation demonstrated how such models can be developed and deployed using widely adopted tools and frameworks, including Python and Visual Studio Code.

We assessed the performance of various predictive algorithms—such as Decision Trees, Random Forest, XGBoost, and Neural Networks—highlighting their respective strengths and limitations. The final application provided a concrete example of how advanced technologies can be leveraged to monitor environmental risks in real time.

Our findings confirm that the integration of AI into air quality monitoring systems enhances both the precision and scalability of environmental assessments. Furthermore, this integration supports more informed decision-making and raises public awareness about air pollution. From a technical standpoint, the project emphasizes the importance of secure web technologies, clean data pipelines, and optimized deployment environments in the development of robust environmental applications.

In conclusion, this project marks a meaningful step toward the digital transformation of environmental monitoring. As cities continue to evolve into smart, interconnected ecosystems, the tools and methodologies developed here can lay the groundwork for future innovations aimed at sustainability, public health protection, and urban resilience. Future work may focus on integrating additional real-time data sources, enhancing model accuracy through hybrid approaches, and deploying the system on scalable cloud infrastructures with advanced security mechanisms. Moreover, the integration of digital twin technology could offer a dynamic virtual representation of the environment, enabling continuous simulation, prediction, and optimization of air quality in real-time urban contexts.

References

[1] World Health Organization. (2021). Air pollution. Retrieved from https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[2] Seinfeld, J. H., Pandis, S. N. (2016). Atmospheric Chemistry and Physics: From Air Pollution to Climate Change. John Wiley Sons.

[3] United Nations Environment Programme. (2019). Air Pollution in Asia and the Pacific: Science-based Solutions. Retrieved from https://www.ccacoalition.org/en/resources/air-pollution-asia-and-pacific-science-based-solutions

[4] U.S. Environmental Protection Agency. (2020). Criteria Air Pollutants. Retrieved from https://www.epa.gov/criteria-air-pollutants

[5] Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., ... Forouzanfar, M. H. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. The Lancet, 389(10082), 1907-1918.

[6] Akimoto, H. (2003). Global air quality and pollution. *Science*, 302(5651), 1716-1719.

[7] Intergovernmental Panel on Climate Change (IPCC). (2021). Climate Change 2021: The Physical Science Basis. Retrieved from https://www.ipcc.ch/report/ar6/wg1/

[8] Holstius, D. M., Pillarisetti, A., Smith, K. R., Seto, E. (2014). Field calibrations of a low-cost aerosol sensor at a regulatory monitoring site in California. *Atmospheric Measurement Techniques*, 7(4), 1121-1131.

[9] European Environment Agency. (2019). Air quality in Europe — 2019 report. Retrieved from https://www.eea.europa.eu/publications/air-quality-in-europe-2019

[10] Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A. (2012). Real-time air quality forecasting, part I: History, techniques, and current status. *Atmospheric Environment*, 60, 632-655.

[11] World Health Organization. (2021). Air quality and health. Retrieved from https://www.who.int/

[12] U.S. Environmental Protection Agency. (2020). Air Quality Index. Retrieved from https://www.epa.gov/

[13] National Aeronautics and Space Administration. (2020). Air pollution and its impacts. Retrieved from https://www.nasa.gov/

[14] U.S. Department of Energy. (2019). Carbon monoxide sources. Retrieved from https://www.energy.gov/

[15] U.S. Environmental Protection Agency. (2020). Volatile Organic Compounds. Retrieved from https://www.epa.gov/

[16] U.S. Environmental Protection Agency. (2020). Ozone formation and its effects. Retrieved from https://www.epa.gov/

[17] National Institute of Environmental Health Sciences. (2020). Particulate matter (PM2.5) and health. Retrieved from https://www.niehs.nih.gov/

[18] World Bank. (2018). The cost of air pollution. Retrieved from https://www.worldbank.org/

[19] World Health Organization. (2021). Air pollution and health. https://www.who.int/

[20] Health Effects Institute. (2019). State of Global Air. https://www.stateofglobalair.org/

[21] The Lancet Neurology. (2020). Air pollution and the brain. https://www.thelancet.com/

[22] U.S. Environmental Protection Agency. (2020). Acid rain. `https://www.epa.gov/`

[23] Intergovernmental Panel on Climate Change. (2021). Climate Change Report. `https://www.ipcc.ch/`

[24] United Nations Environment Programme. (2020). Effects of air pollution on biodiversity. `https://www.unep.org/`

[25] World Bank. (2016). The Cost of Air Pollution. `https://www.worldbank.org/`

[26] Organisation for Economic Co-operation and Development. (2014). The cost of air pollution. `https://www.oecd.org/`

[27] Food and Agriculture Organization. (2018). Air pollution and agriculture. `https://www.fao.org/`
99

[28] NASA. (2020). MODIS Air Quality Data. `https://modis.gsfc.nasa.gov/`

[29] European Space Agency. (2021). Sentinel-5P Data. `https://sentinels.copernicus.eu/`

[30] U.S. Environmental Protection Agency. (2021). Air Quality Index (AQI). `https://www.airnow.gov/`

[31] World Health Organization. (2021). WHO global air quality guidelines. `https://www.who.int/publications/i/item/9789240034228`

[32] U.S. Environmental Protection Agency. (2021). National Ambient Air Quality Standards. `https://www.epa.gov/naaqs`

[33] European Union. (2008). Directive 2008/50/EC. `https://eur-lex.europa.eu/`

[34] Kumar, P. et al. (2015). The rise of low-cost sensing for managing air pollution in cities. *Environment International*, 75, 199-205.

[35] Castell, N. et al. (2017). Can commercial low-cost sensor platforms contribute to air quality monitoring? *Environment International*, 99, 293-302.

[36] UCI Machine Learning Repository. (2020). Beijing Multi-site Air-Quality Data. `https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data`

[37] Kaggle. (2021). Air Quality Datasets. `https://www.kaggle.com/datasets`

[38] OpenAQ. (2020). Open Air Quality Data Platform. `https://openaq.org/`

[39] Snyder, E. G. et al. (2013). The changing paradigm of air pollution monitoring. *Environmental Science & Technology*, 47(20), 11369–11377.

[40] Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A. (2012). Real-time air quality forecasting, part I: History, techniques, and current status. *Atmospheric Environment*, 60, 632–655.

[41] Chakma, A., Bhattacharyya, D. (2022). Air quality prediction using machine learning: A comprehensive review. *Environmental Science and Pollution Research*, 29, 29568–29595.

[42] Gonzalez, M., Valera, H. G., Rodriguez, M. (2020). Predicting air pollution with machine learning: A case study in Madrid. *Journal of Cleaner Production*, 256, 120371.

[43] Cheng, X., Chen, T. (2021). IoT-enabled air quality monitoring and forecasting: A review. *IEEE Internet of Things Journal*, 8(17), 13407–13423.

[44] Cheng, Y., Zhang, L., Zhou, M. (2021). A review of air pollution prediction using machine learning methods. *Environmental Science and Pollution Research*, 28(8), 9575–9594.

[45] Zhou, Y., Chang, H. H. (2019). Application of Ridge Regression in Air Quality Prediction. *Atmospheric Pollution Research*, 10(5), 1513–1520.

[46] Li, X., Peng, L., Hu, Y. (2020). Classification of air pollution levels using Random Forest. *Environmental Modelling Software*, 124, 104601.

[47] Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

[48] Jiang, D., Wang, F., Ma, X. (2021). Unsupervised learning in air pollution pattern detection using K-Means and DBSCAN. *Environmental Informatics*, 45, 67–74.

[49] Liu, H., Yin, H., Wang, Y. (2019). Semi-supervised learning for AQI prediction: A hybrid framework. *IEEE Access*, 7, 75697–75709.

[50] Sun, W., Xu, J., Ma, X. (2020). PCA-based feature selection for air pollution prediction. *Environmental Monitoring and Assessment*, 192(6), 392.

[51] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? *Geoscientific Model Development*, 7(1), 1247–1250.

[52] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.

[53] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? *Geoscientific Model Development*, 7(1), 1247–1250.

[54] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.

[55] Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A. (2018). Real-time air quality forecasting, Part I: History, techniques, and current status. *Atmospheric Environment*, 199, 336-350.

[56] Hagenauer, J., Helbich, M. (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications*, 78, 273–282.

[57] Cheng, J., Liu, F., Yao, Y. (2018). Air quality prediction method based on LSTM neural network. *IOP Conf. Series: Earth and Environmental Science*, 113(1), 012127.

[58] Wang, Y., Gu, X., Zhang, C., Yu, Z. (2021). A hybrid machine learning model for PM2.5 prediction using empirical mode decomposition and multiple deep learning models. *Science of the Total Environment*, 758, 143564.

[59] Zheng, Y., Liu, F., Hsieh, H. P. (2013). U-Air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD*.

[60] Guidotti, R., et al. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 93.

[61] Chen, L., et al. (2021). Transfer learning for air quality prediction: A case study in China. *Environmental Science and Pollution Research*, 28(9), 10987–11000.

[62] Ghahramani, Z., et al. (2021). Smart city architecture for real-time air quality monitoring using AI. *IEEE Internet of Things Journal*.

[63] Lu, Y. (2022). Ethics of artificial intelligence in environmental monitoring: A review. *AI and Society*.

[64] Python Software Foundation. (2023). *Official Python Documentation.* Retrieved from `https://www.python.org/doc/`

[65] VanderPlas, J. (2016). *Python Data Science Handbook.* O'Reilly Media.

[66] Lutz, M. (2013). *Learning Python* (5th ed.). O'Reilly Media.

[67] McKinney, W. (2012). *Python for Data Analysis.* O'Reilly Media.

[68] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

[69] Streamlit. (2023). *Streamlit Documentation*. Retrieved from `https://docs.streamlit.io/`

[70] Treuille, A., Snow, J. (2021). *The Streamlit Guide: Building Data Apps with Python*. Streamlit Inc.

[71] The pandas development team. (2023). *pandas documentation*. Retrieved from `https://pandas.pydata.org/`

[72] Harris, C. R., et al. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362.

[73] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.