SMART AQI FORECASTING AND PUBLIC AIR PURIFIER DEPLOYMENT FOR POLLUTION MITIGATION

PRATIK PRAKASH PATIL

Final Thesis Report

MAY 2025

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Full Form
AQI	Air Quality Index
IoT	Internet of Things
ML	Machine Learning
LSTM	Long Short-Term Memory
PM2.5	Particulate Matter ≤ 2.5 microns
PM10	Particulate Matter ≤ 10 microns
NO_2	Nitrogen Dioxide
SO_2	Sulfur Dioxide
CO	Carbon Monoxide
O_3	Ozone
RF	Random Forest

CHAPTER 1: INTRODUCTION

1. Introduction

Air pollution remains a critical global challenge, significantly impacting public health and environmental sustainability. Smart Air Quality Index (AQI) forecasting and strategic public air purifier deployment are emerging as effective solutions for pollution mitigation. Traditional air quality monitoring systems rely on static sensors, which often fail to provide real-time, location-specific insights. The amalgamation of internet of Things (IoT), machine learning (ML), as well as sophisticated analytics of data can enhance AQI forecasting accuracy, enabling proactive pollution control measures.

Air pollution is a growing environmental and public health concern, particularly in urban areas where industrial activities, vehicular emissions, and construction contribute significantly to deteriorating air quality. The Air Quality Index (AQI) serves as a standardized tool to measure and communicate the level of air pollutants, including PM2.5, PM10, NO_2, SO_2, CO, and O_3. In recent years, the need for accurate AQI forecasting has become increasingly critical for timely interventions and public awareness. Traditional reactive approaches to pollution management often fall short in minimizing exposure to harmful air pollutants, especially during peak pollution periods. As a result, there is a pressing demand for intelligent systems that can not only predict AQI levels with high accuracy but also support proactive decision-making in pollution mitigation.

The concept of "Smart AQI Forecasting and Public Air Purifier Deployment" addresses this challenge by integrating data analytics, machine learning, and IoT-enabled infrastructure. This system utilizes real-time and historical environmental data to forecast AQI levels with predictive accuracy, allowing authorities to take early action. Advanced models such as LSTM, Random Forest, and hybrid neural networks are employed to analyse complex air quality patterns and meteorological influences. Coupled with this forecasting capability is the strategic deployment of public air purifiers in high-risk zones. These purifiers, equipped with smart sensors and connectivity features, can dynamically respond to pollution levels and operate efficiently where needed most.

This research aims to develop a smart AQI forecasting model that leverages deep learning techniques, real-time sensor data, and meteorological parameters to predict air pollution levels with high precision. Additionally, an optimized public air purifier deployment strategy will be designed using spatial analysis and predictive modelling to mitigate pollution in high-risk areas. The proposed framework will dynamically adjust purifier placements based on pollution hotspots, traffic density, and weather conditions.

By combining intelligent AQI prediction with adaptive purification strategies, this study seeks to improve urban air quality, reduce health risks, and optimize resource utilization. The findings will contribute to the development of sustainable, data-driven pollution management systems, supporting smart city initiatives and enhancing public well-being.
Researchers have developed various machine learning models to predict particulate matter (PM) concentrations. Techniques such as linear regression, random forest, K-nearest neighbors (KNN), ridge and lasso regression, XG Boost, and AdaBoost have been employed to forecast PM levels, enhancing the accuracy of AQI predictions. Innovations include hybrid deep learning models that combine Attention Convolutional Neural Networks (ACNN) with Autoregressive Integrated Moving Average (ARIMA) models. These approaches have improved the precision of AQI forecasts by capturing both spatial and temporal patterns in air quality data. Studies have compared different regression techniques—such as Random Forest, Linear Regression, & Decision Tree Regression—to identify most effective models for air quality prediction in urban settings. Findings suggest that Decision Tree Regression often exceeds the efficiency of alternative models in terms of accuracy and computational efficiency.

Recent developments include AI-driven public air purifying systems that integrate sensors (e.g., MQ135 gas sensors), water spray mechanisms, and HEPA filters. These systems dynamically monitor and improve air quality in real-time, adjusting their operations based on pollution levels to optimize performance and energy consumption. Projects like "Smart Breathe" have focused on deploying IoT-integrated air purification units across high-density urban areas. These interconnected units continuously monitor environmental conditions and adapt their purification strategies accordingly, leading to localized improvements in air quality and public health. Cities have introduced clean air zones to reduce pollution from vehicular emissions. For instance, Bradford's clean air zone has led to significant reductions in air pollution and healthcare savings,

demonstrating the effectiveness of policy measures in conjunction with technological solutions.

1.1 Background of the Study

Air pollution has been a growing concern since the Industrial Revolution, with rapid urbanization and industrialization significantly deteriorating air quality in cities worldwide. Historically, efforts to manage pollution relied on reactive measures, such as warning systems and emissions regulations, which often fell short in preventing health risks associated with poor air quality. Over time, the development of Air Quality Index (AQI) systems and the integration of environmental monitoring technologies have improved our ability to measure and interpret pollution levels. Recent advancements in machine learning, IoT, and big data analytics have opened new possibilities for real-time air quality forecasting, enabling proactive mitigation strategies.

Despite these technological strides, existing systems often lack the predictive precision and actionable integration required for timely public health interventions. Many forecasting models either provide short-term estimates or operate with limited spatial granularity, making them insufficient for dynamic urban environments. Additionally, the deployment of public air purifiers has largely been reactive and uncoordinated, with little data-driven strategy behind their placement or effectiveness.

This study aims to bridge these gaps by developing a smart AQI forecasting model integrated with a strategic framework for real-time public air purifier deployment. By combining predictive analytics with spatial optimization, the proposed system can enhance urban resilience to pollution, reduce exposure risks, and support evidence-based policy-making for healthier cities.

1.2 Problem Statement

Pollution of the environment represents a critical challenge to the well-being of the population as well as environmental sustainability, necessitating efficient monitoring and mitigation strategies. Traditional Air Quality Index (AQI) forecasting methods often lack accuracy, real-time adaptability, and predictive capabilities, making it difficult to implement

timely interventions. Additionally, the deployment of public air purifiers is not optimized, leading to inefficient resource utilization and limited impact on pollution reduction.

This study aims to develop a Smart AQI Forecasting System using advanced machine learning techniques to provide precise, real-time air quality predictions. Furthermore, it focuses on optimizing the strategic deployment of public air purifiers based on pollution patterns, population density, and meteorological factors. By integrating AI-driven forecasting with intelligent air purifier placement, this research seeks to enhance urban air quality management, minimize health risks, and improve environmental sustainability.

1.3 Aim and Objectives

The aim of this study is to develop an intelligent Air Quality Index (AQI) forecasting system integrated with an optimized public air purifier deployment strategy to mitigate pollution in urban environments. By leveraging advanced ML techniques, actual-time sensor information, and predictive analytics, the system aims to deliver precise both short- as well as long-term AQI forecasts. This information will facilitate proactive pollution management and assist in strategically deploying public air purifiers in high-risk areas. The ultimate goal is to enhance air quality, reduce human exposure to pollutants, and support smart city initiatives by enabling data-driven environmental decision-making and resource optimization.

The research objectives are derived from this aim and include:

1) Develop advanced ML as well as deep learning techniques to determine Air Quality Index (AQI) with high accuracy.
2) Utilize IoT-enabled air quality sensors to collect actual-time pollution information through different locations.
3) Strategically deploy smart air purifiers in high-pollution zones based on AQI forecasts and real-time air quality data.
4) Optimize the operation of air purifiers by adjusting their intensity and placement dynamically based on pollution levels.
5) Analyze historical AQI trends and real-time data to identify major pollution sources and contribute to targeted mitigation strategies.

1.4 Research Questions

Which machine learning model yields the highest accuracy in predicting Air Quality Index (AQI) for different cities, and how can this information be utilized to strategically deploy air purifiers in these cities to improve air quality and public health?

1.5 Scope of the Study

The scope of this study encompasses the development of a Smart Air Quality Index (AQI) Forecasting System integrated with Public Air Purifier Deployment to mitigate pollution in urban environments. With the increasing concerns over air pollution and its adverse effects on public health, there is a growing need for advanced solutions that provide real-time AQI monitoring and proactive mitigation strategies. This study aims to leverage machine learning (ML) and deep learning (DL) models for accurate air quality predictions based on meteorological data, pollutant concentrations, and traffic emissions. By incorporating past & actual-time AQI information, the forecasting model can help authorities make informed decisions on pollution control measures.

Furthermore, the study explores the optimal deployment of public air purifiers in highly polluted areas based on predictive analysis. Instead of random placement, a data-driven approach will be used to identify pollution hotspots and determine the most effective locations for purifier installation. The study also considers factors such as population density, wind patterns, and industrial emissions to maximize efficiency.

A key aspect of this research is the incorporation of IoT sensors and cloud-based AI systems to facilitate continuous monitoring and adaptive control of air purifiers. This will ensure real-time adjustments in purification intensity based on pollution severity, thereby improving air quality dynamically.

The findings of this study will be valuable for urban planners, policymakers, and environmental agencies, enabling the creation of smart, pollution-resilient cities. Additionally, the research opens avenues for future enhancements, such as AI-driven autonomous purification networks, improved sensor accuracy, and citizen engagement

through mobile applications. By implementing this intelligent air quality management system, cities can take a significant step toward sustainable environmental solutions and public health protection.

1.6 Significance of the Study

The significance of this study lies in its comprehensive use of machine learning (ML) and statistical techniques to predict the air quality index (AQI), which is crucial for public health and environmental management. Key points of significance include:

- **Improved Air Quality Monitoring** – The study enhances real-time air quality index (AQI) forecasting using smart technologies, providing accurate pollution level predictions.
- **Data-Driven Decision Making** – AI and machine learning-based forecasting enable governments and policymakers to make informed decisions on pollution control strategies.
- **Targeted Public Air Purifier Deployment** – Optimized placement of public air purifiers ensures efficient pollution reduction in high-risk areas, improving urban air quality.
- **Health Benefits** – Reducing air pollution exposure lowers respiratory diseases and cardiovascular risks, leading to overall public health improvements.
- **Sustainable Urban Planning** – The study supports smart city initiatives by integrating environmental intelligence with urban infrastructure for long-term sustainability.
- **Cost-Effective Pollution Control** – Proactive AQI forecasting and strategic air purifier deployment reduce the financial burden of large-scale pollution mitigation measures.
- **Public Awareness and Engagement** – Real-time AQI updates empower citizens to take protective measures, promoting eco-friendly behaviours and community participation.
- **Technological Advancements** – Encourages further innovation in IoT, AI, and big data analytics for air quality monitoring and environmental management.

1.7 Structure of the Study

Chapter 1: Introduction

Introduces the research problem, significance, and objectives. It outlines the impact of air pollution, the need for accurate AQI forecasting, and the role of public air purifier systems in urban pollution mitigation.

Chapter 2: Literature Review

Surveys existing work on air quality monitoring, AQI prediction techniques (e.g., machine learning, time series models), and air purification technologies. Identifies research gaps in predictive accuracy and real-time purifier deployment.

Chapter 3: Methodology

Describes the data collection process, model selection, and forecasting framework. It details the use of smart algorithms (e.g., LSTM, Random Forest, XG boost, Kalman filter) for AQI prediction and optimization strategies for deploying air purifiers based on predicted pollution hotspots.

Chapter 4: System Design and Implementation

Explains the development of the smart AQI forecasting system, sensor network integration, and the design of a real-time purifier control system. Includes user interface elements and backend architecture.

Chapter 5: Results and Discussion

Presents experimental results, model evaluation metrics, and comparative performance analysis. Discusses the system's effectiveness in accurately forecasting AQI and the practical benefits of dynamic purifier deployment.

Chapter 6: Conclusion and Future Work

Summarizes key findings, contributions, and system limitations. Proposes future enhancements such as incorporating weather data, expanding sensor networks, and integrating citizen feedback.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Air pollution has emerged as one of the most critical environmental and public health challenges of the 21st century, with rising urbanization and industrialization contributing significantly to deteriorating air quality worldwide. Accurate forecasting of Air Quality Index (AQI) has become essential for timely mitigation strategies, enabling authorities to issue health advisories and take preventive actions. Traditional AQI monitoring systems, while useful, often suffer from limited spatial coverage and delayed reporting, which restrict their effectiveness in real-time decision-making. In response to these limitations, smart AQI forecasting using machine learning, artificial intelligence (AI), and Internet of Things (IoT) technologies has gained increasing attention in recent years. These intelligent systems can predict pollution levels based on dynamic environmental parameters, traffic data, and historical air quality trends, offering more precise and proactive pollution management tools.

One promising strategy for pollution mitigation is the deployment of public air purifiers, which act as immediate intervention tools to reduce localized particulate matter concentrations in high-risk zones. Integrating AQI forecasting with the strategic placement of public air purifiers creates a responsive, data-driven approach to managing urban air quality. Recent literature explores models that optimize purifier placement based on real-time AQI predictions, population density, and pollution sources. Moreover, research emphasizes the importance of combining technological innovation with public policy, community awareness, and sustainable urban planning to ensure long-term effectiveness. This review examines the evolution of smart AQI forecasting methods, including deep learning models, hybrid sensor networks, and geospatial data analytics, as well as the current advancements in air purifier technologies and deployment strategies. It highlights the potential of integrated systems that utilize smart forecasting to guide the efficient use of air purification resources, ultimately aiming to minimize pollution exposure and protect public health in urban environments.

Anikender Kumar and Pramila Goyal (2011) conducted a study to predict the daily Air Quality Index (AQI) for Delhi using historical AQI records and meteorological data. They applied Principal Component Regression (PCR) and Multiple Linear Regression

(MLR) models, training on data from 2000 to 2005 and testing on AQI values from 2006. PCR was specifically employed to address multicollinearity among input variables by reducing dimensionality, and the resulting components were fed into MLR for AQI prediction across different seasons. The model showed highest accuracy during winter. However, the study only used meteorological variables and did not include pollutant concentrations, which are key contributors to health-related air quality outcomes.

Huixiang Liu et al. (2019) analyzed air quality data from Beijing and an Italian city. They used Support Vector Regression (SVR) and Random Forest Regression (RFR) to predict AQI in Beijing and nitrogen oxides (NOx) concentrations in Italy using two distinct datasets. The Beijing dataset (2013–2018) included hourly pollutant levels (PM2.5, PM10, SO2, O3, NO2), while the Italian dataset (2004–2005) focused on NOx-related pollutants. SVR outperformed RFR for AQI prediction, while RFR was better suited for NOx estimation.

Ziyue Guan and Richard O. Sinnott (2018) leveraged machine learning models—Artificial Neural Networks (ANN), Linear Regression (LR), and Long Short-Term Memory (LSTM)—to forecast PM2.5 levels in Melbourne. The study utilized both official EPA data and unofficial PM2.5 measurements from Airbeam devices. Among the tested models, LSTM exhibited superior predictive performance and was effective in capturing extreme PM2.5 values.

Heidar Maleki et al. (2019) used Artificial Neural Networks (ANN) to forecast hourly concentrations of pollutants (NO2, SO2, PM10, PM2.5, CO, O3) and to compute both AQI and Air Quality Health Index (AQHI) for four monitoring sites in Ahvaz, Iran. Their models incorporated pollutant levels, meteorological conditions, and time-based features using data collected from August 2009 to August 2010.

Aditya C. R. et al. (2018) implemented logistic regression and auto-regression to classify pollution levels and forecast PM2.5 concentrations. Their dataset consisted of atmospheric variables from a specific city, and the dual-model approach aimed to first detect pollution and then predict future pollutant values.

Nidhi Sharma et al. (2018) conducted a time-series analysis on air quality data in Delhi (2009–2017), emphasizing pollution trends from 2016–2017. Using time series regression

models, they forecasted future levels of SO2, NO2, PM10, PM2.5, CO, Benzene, and O3. Their study revealed increasing trends in PM10, NO2, and PM2.5 levels, while CO and Benzene were projected to decrease slightly.

Mohamed Shakir and N. Rakesh (2018) examined diurnal and weekly patterns of air pollutants in Karnataka using WEKA tools. Their analysis used ZeroR and K-means clustering to identify how pollutants like NO, NO2, CO, PM10, and SO2 varied with temperature, humidity, and wind speed. Results suggested peak pollution levels during workdays and business hours.

Kazem Naddaf et al. (2012) applied the WHO's AirQ model to assess the health impact of air pollutants (PM10, SO2, NO2, and O3) in Tehran. Their findings showed PM10 had the most significant effect on public health, with an estimated 2,194 excess deaths annually attributable to its presence. The study underlined the urgent need to reduce PM10 concentrations to mitigate health risks.

Yusef Omidi Khaniabadi et al. (2016) also used AirQ to evaluate the connection between pollutant levels and cardiovascular mortality in Kermanshah, Iran. Their results indicated a clear correlation, estimating that a 10 $\mu g/m^3$ increase in PM10, NO2, or O3 significantly raises mortality risk.

R. Gunasekaran et al. (2012) monitored air quality around Salem Swadeswari College, Tamil Nadu, from April 2010 to March 2011. While most pollutants were within national limits, PM10 levels exceeded the 24-hour average standard for most months, suggesting moderate air quality concern.

S. Tikhe Shruti et al. (2013) used Artificial Neural Networks (ANN) and Genetic Programming (GP) to model and predict pollutant levels (SOx, NOx, RSPM) in Pune between 2005 and 2011. Their results showed that GP models outperformed ANN in forecasting accuracy, suggesting the potential of evolutionary algorithms in environmental modeling.

R. Sharma, G. Shilimkar, and S. Pisal (2021) conducted a comprehensive study on air quality prediction using machine learning techniques, focusing on the application of supervised learning models to forecast AQI levels based on historical environmental data.

Their work highlights the effectiveness of algorithms such as decision trees, random forest, and support vector machines (SVM) in capturing complex patterns in pollutant data, including PM2.5, PM10, NO_2, and SO_2. The researchers emphasize data preprocessing steps like normalization and feature selection to improve model accuracy and reduce noise. Their experiments demonstrated that ensemble methods like random forest performed better in terms of accuracy and robustness compared to single algorithms. The study concludes that machine learning offers a promising approach for real-time AQI forecasting, which can be instrumental for early warning systems and pollution control strategies. It also underscores the potential for integrating such models with smart city infrastructure to support data-driven environmental policy decisions.

Zhang, Chen, and Huang (2023) presented a novel approach for air quality prediction by integrating wavelet transform, Detrended Cross-Correlation Analysis (DCCA), and a Long Short-Term Memory (LSTM) model. Their study, published in Applied Sciences, focuses on improving the accuracy of AQI forecasting by capturing both the temporal dynamics and the long-range dependencies inherent in air pollution data. The wavelet transform was used to decompose the time series data into multiple frequency components, enabling the model to analyse patterns at various scales. The DCCA technique further enhanced the model by identifying significant correlations between pollutant variables over time. Finally, the LSTM network leveraged these processed inputs to forecast AQI with greater precision. The results demonstrated that the hybrid model outperformed conventional LSTM and other baseline models in terms of prediction accuracy, especially in scenarios involving complex pollutant interactions and fluctuating environmental conditions. This study highlights the effectiveness of combining signal processing techniques with deep learning for robust air quality forecasting.

2.3 Summary

Recent research on air quality forecasting and pollution mitigation highlights the growing use of machine learning (ML) and data analytics in predicting AQI and deploying interventions. Kumar and Goyal (2011) applied Principal Component Regression (PCR) and Multiple Linear Regression to forecast AQI in Delhi, finding PCR effective, especially in winter. Liu et al. (2019) used Support Vector Regression and Random Forest to predict AQI in Beijing and NOx levels in Italy, with SVR excelling in AQI forecasting. Guan and Sinnot (2018) applied LSTM networks, achieving accurate PM2.5 predictions using official and sensor data. Maleki et al. (2019) predicted hourly pollutants and AQI in Ahvaz using ANN, while Aditya et al. (2018) combined Logistic and Auto Regression for PM2.5 forecasting. Sharma et al. (2021) confirmed the strength of ensemble ML models like random forest for AQI prediction. Zhang et al. (2023) introduced a hybrid wavelet-LSTM model with DCCA for enhanced AQI forecasting. Studies like those by Khaniabadi and Naddaf used WHO's AirQ tool to quantify health impacts of pollution in Iran. Collectively, these studies show that integrating ML with pollutant data and meteorological parameters supports smarter AQI forecasting and effective public air purifier deployment strategies.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

Air pollution remains one of the most pressing environmental and public health challenges of the 21st century, especially in rapidly urbanizing regions. Traditional reactive approaches to pollution control are often inadequate in preventing exposure to harmful air quality levels. To address this, smart AQI (Air Quality Index) forecasting combined with strategic deployment of public air purifiers offers a forward-looking solution for pollution mitigation. By leveraging advanced machine learning models, real-time environmental data, and predictive analytics, it is now possible to anticipate pollution surges with high accuracy. These forecasts enable timely decisions, such as activating or repositioning public air purifiers in high-risk areas like schools, hospitals, and urban hotspots. Integrating technology with public health infrastructure not only enhances community resilience but also promotes awareness and proactive engagement. This innovative approach represents a shift toward data-driven, anticipatory environmental management, aiming for cleaner air and healthier living conditions in densely populated areas.

3.2 Research Methodology

The research methodology for the study titled "Smart AQI Forecasting and Public Air Purifier Deployment for Pollution Mitigation" integrates data-driven approaches and system design to address urban air quality challenges. This study adopts a mixed-methods framework, combining quantitative data analysis, predictive modelling, and geospatial optimization techniques. The methodology begins with the collection of historical air quality data, meteorological variables, and pollutant concentration levels from publicly available sources and real-time sensors. Machine learning algorithms, such as Random Forest, LSTM, or Gradient Boosting, are then employed to develop a predictive model for forecasting AQI (Air Quality Index) across different urban zones. The forecasting model is trained and validated using statistical performance metrics such as RMSE, MAE, and R^2 to ensure accuracy and reliability. Based on forecasted pollution hotspots, a geospatial analysis is conducted using tools like GIS or optimization algorithms to strategically determine optimal locations for deploying public air purifiers. Additionally, simulation

scenarios are used to evaluate the effectiveness of purifier deployment in reducing local pollutant concentrations. The research also considers socio-environmental parameters to ensure equitable distribution and accessibility. This methodological approach provides a comprehensive framework to forecast pollution levels and deploy mitigation infrastructure intelligently, supporting smart city initiatives and public health protection.

3.2.1 Data Selection

Data selection is the process of identifying and extracting relevant and high-quality data from various sources to address a specific problem or objective. In the context of projects like AQI forecasting or pollution mitigation, it involves choosing the most appropriate datasets—such as air quality measurements, weather conditions, traffic patterns, and geographic information—that directly influence or reflect the factors being studied. Effective data selection ensures that the chosen data is accurate, representative, timely, and sufficient for analysis. This step is critical for building reliable predictive models, optimizing resources, and making informed decisions. Poor or irrelevant data can lead to misleading results and ineffective solutions. Therefore, careful selection based on criteria like relevance, completeness, accuracy, and consistency is essential to achieve meaningful outcomes in data-driven projects.

To ensure data quality and relevance, the following procedures were incorporated during the data selection and preparation phase:

I. **Duplicate Removal**: Any exact duplicate entries were removed to prevent bias or distortion in the model due to repeated values.

II. **Missing Value Handling**: Columns with an excessive number of missing values were excluded, and for the rest, imputation strategies were used to fill gaps where necessary.

III. **Outlier Detection and Treatment**: The IQR (Interquartile Range) method was employed to detect and cap outliers in key pollutant variables (e.g., PM2.5, NO_2, CO, AQI) to reduce skewness and enhance model robustness.

IV. **Datetime Cleaning and Feature Engineering**: The 'Datetime' column was converted to a proper datetime format and decomposed into cyclic and categorical features (Hour, Weekday, Month, Season), enabling time-series modeling and seasonal trend detection.

V. **Categorical Encoding**:

a. One-Hot Encoding was used for categorical time features like ‘Weekday’ and ‘Season’ to avoid ordinality assumptions.

b. Label Encoding was applied to ordinal categorical features such as the AQI bucket ‘Status’ for classification tasks.

VI. **Irrelevant/Constant Columns Removal**: Non-informative or constant columns (e.g., blank, null, or repeated values) were dropped to reduce dimensionality and improve processing efficiency.

VII. **Final Data Reduction**: After cleaning, the dataset was reduced to a manageable size of **902,461 rows and 30 columns**, retaining only the most informative and relevant attributes for modeling AQI levels. These cleaning and preprocessing steps ensure the selected dataset is fit for building high-performing models for both AQI forecasting and classification tasks.

1. Historical and Real-Time AQI Data

- Sources: Government agencies (e.g., EPA, CPCB), OpenAQ, WAQI, local environmental sensors
- Parameters: PM2.5, PM10, NO_2, SO_2, CO, O_3, AQI Index values
- Granularity: Hourly/daily readings by location

2. Meteorological Data

- Sources: Weather APIs (OpenWeatherMap, NOAA, IMD)
- Parameters: Temperature, humidity, wind speed/direction, precipitation, solar radiation

- Use: Enhances AQI forecasting models by understanding pollutant dispersion

3. Geospatial and Urban Data

- GIS Layers: Population density, land use, traffic density, green cover
- Sources: OpenStreetMap, satellite imagery, urban planning databases
- Use: To optimize purifier placement in high-risk or high-footfall areas

4. Traffic and Mobility Data

- Sources: Traffic sensors, Google Maps APIs, city traffic departments
- Parameters: Vehicle count, traffic speed, congestion levels
- Use: To correlate with pollution hotspots and real-time AQI spikes

5. Public Health and Exposure Data

- Sources: Hospital reports, health surveys, asthma/emergency visits
- Use: For prioritizing deployment in sensitive zones like schools and hospitals

6. Purifier Deployment Data

- Parameters: Location, purifier type, capacity, operational status
- Use: For real-time performance tracking and adaptive deployment

3.2.2 Data Pre-processing

Data preprocessing is the process of cleaning, transforming, and organizing raw data into a suitable format for analysis or modelling. It involves steps such as handling missing values, removing duplicates, normalizing or scaling features, encoding categorical variables, and detecting outliers. This step is essential to improve data quality, enhance model accuracy, and ensure that machine learning algorithms can effectively learn from the data. Without proper preprocessing, models may produce unreliable or biased results.

In order to create precise time series forecasting models, the quality of the integrity of the time sequence information is crucial. It has a direct effect on parameter estimation accuracy and model performance. Missing values, unstructured timestamps, and outliers present serious problems when it comes to air quality data. These problems, which result in imprecise or insufficient air quality readings, can be caused by a number of things, such as malfunctioning monitoring stations or outside influences.

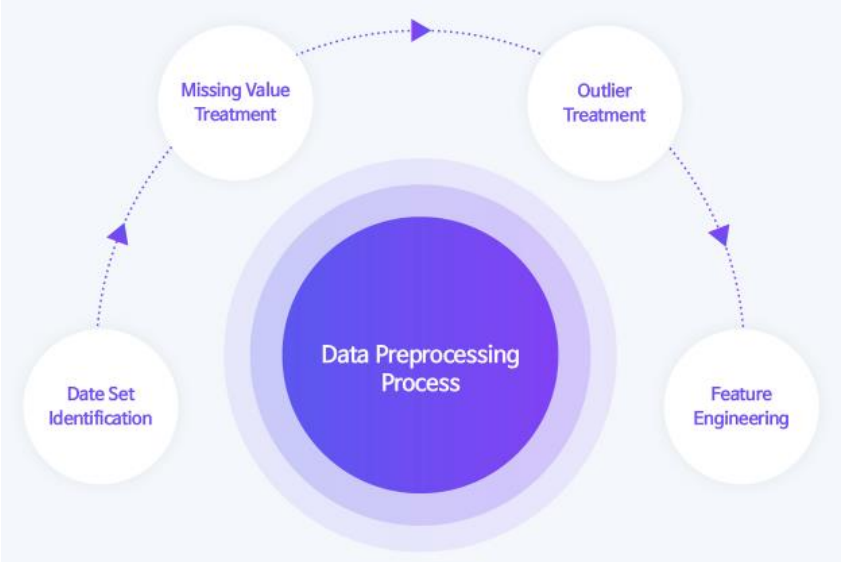

Figure 3.1 : Data Pre-Processing

The following pre-processing steps were used when developing a hybrid deep learning model:

3.2.2.1 Data Collection

1. Sources: Gathered data from multiple repositories including station_hour.csv and stations.csv, which contain IoT-based air quality readings across monitoring locations in India. Supplementary sources may include meteorological stations, satellite imagery, and government AQI databases like CPCB and OpenAQ.
2. Parameters: Included pollutant concentrations (PM2.5, PM10, NO_2, SO_2, CO, O_3), meteorological variables (humidity, temperature, wind speed), and geospatial station metadata (latitude, longitude, location category).

3.2.2.2 Data Cleaning

1. Handling Missing Data: Columns with excessive missing values were dropped. For other features, missing values were handled using statistical imputation techniques such as mean and median imputation.
2. Outlier Detection and Treatment: Applied the **IQR method** to cap outliers in key pollutant and AQI variables, enhancing the distribution of the data and improving model robustness.
3. Duplicate Removal: Removed duplicate records to prevent data redundancy and inflated influence during training. Removed duplicate records to prevent skewing the model results.
4. Datetime Formatting: Standardized the Datetime column, which was initially in an inconsistent format, ensuring consistency for further temporal feature extraction.
5. Noise Reduction: Apply moving averages or Kalman filtering to smooth fluctuating sensor data.

3.2.2.3 Feature Engineering and Transformation

1. Normalization: Scale features using Min-Max or Z-score normalization for machine learning models.
2. Encoding Categorical Data: Convert categorical features (e.g., location, weather conditions) utilizing one-hot encoding either label encoding.
3. One-Hot Encoding was applied to time-based features such as Weekday and Season, transforming them into binary variables.
4. Label Encoding was used for the AQI Status column (with values like Good, Moderate, etc.), appropriate for classification tasks.
5. Feature Engineering: Extracted time-related features from the Datetime column:
    1. Hour, Day, Month, Weekday, Season
    2. Applied cyclical encoding (sine and cosine transformations) for features like Hour and Month to handle the cyclical nature of time.

6. Utility Engineering: Generate new utility such as AQI trends, pollutant ratio indicators, and historical lag features.

7. Column Cleanup: Removed any irrelevant columns, like empty or constant-value fields, ensuring the dataset focused only on valuable features.

3.2.2.4 Time-Series Processing

- Resampling: Convert raw sensor readings into uniform time intervals (e.g., hourly/daily averages).
- Granularity: Maintained the hourly resolution as provided in the dataset, enabling accurate prediction of AQI trends at different times of the day and season.
- Sliding Window Method: Create time-based features for better AQI forecasting using past observations.
- Temporal Feature Extraction: Engineered time-based features to better capture temporal patterns such as daily and seasonal variations in pollution levels.

3.2.2.5 Data Integration

- Merging Multiple Data Sources: Integrated station_hour.csv and stations.csv on the StationId, combining air quality and pollutant levels with geographical station metadata, creating a robust dataset with over 2.5 million records. Combine real-time AQI sensor data with weather forecasts and historical pollution trends.
- Spatial Data Mapping: Combined spatial features (like station location and type) with temporal features (like hour, weekday, season) to enable geographically and time-sensitive AQI forecasting. Use GIS-based preprocessing for optimizing purifier deployment based on high-risk zones.

3.2.2.6 Splitting Data for Model Training

- Train-Test Split: Typically, an 80-20 or 70-30 split is used for predictive modeling. The cleaned dataset of **902,461 rows** was split for model training, typically using an **80-20 or 70-30 split** for machine learning models.
- Cross-Validation: Implement k-fold cross-validation for robust model evaluation. K-fold cross-validation was considered for model evaluation

ensure robustness and reduce overfitting, especially when dealing with time-series data.

Preprocessed data ensures better AQI forecasting accuracy and optimizes public air purifier placement for effective pollution mitigation.

3.2.3 Data Transformation

Data transformation is the process of converting raw data into a format that is suitable for analysis, modeling, or further processing. It involves applying various techniques to clean, restructure, and standardize data to ensure consistency and accuracy. Common transformation steps include normalization (scaling values), encoding categorical variables, handling missing values, aggregating data, and converting data types. In the context of machine learning or forecasting, such as AQI prediction, transformation helps improve model performance by ensuring all input features are comparable and meaningful. For example, temperature and AQI values may be on different scales, so normalization ensures one doesn't dominate the model. Data transformation also aids in uncovering hidden patterns and relationships, making it a critical step in any data-driven project.

3.2.4 Data Mining

Data mining is the process of discovering meaningful patterns, trends, and relationships within large sets of data using statistical, machine learning, and computational techniques. It transforms raw data into useful information by identifying hidden correlations, anomalies, and predictive insights that are not immediately obvious. Common tasks in data mining include classification, clustering, regression, association rule mining, and anomaly detection. It is widely applied in fields like marketing, finance, healthcare, and environmental science to support decision-making and strategic planning. Data mining typically involves data cleaning, integration, selection, and transformation, followed by the application of analytical models. The ultimate goal is to extract knowledge that adds value, helps forecast

future trends, or explains existing patterns. Tools and techniques used in data mining are often part of broader disciplines like data science and artificial intelligence.

3.2.5 Interpretation/Evaluation

Interpretation and evaluation are critical stages in research and data analysis that help derive meaningful insights and assess the value of results.

Interpretation involves making sense of the data by explaining patterns, trends, and relationships observed. It connects the raw outcomes to the research objectives or hypotheses, helping to understand what the results signify in context. For example, if AQI levels spike during high traffic hours, interpretation would highlight traffic's role in air pollution.

Evaluation assesses the reliability, validity, and significance of the findings. It considers whether the results meet expectations, the limitations of the methodology, and how well the objectives were achieved. In applied projects like AQI forecasting, evaluation also involves judging model accuracy, environmental impact, and effectiveness of interventions like purifier deployment.Together, interpretation and evaluation ensure that conclusions are well-grounded, actionable, and useful for decision-making or further research.

3.3 Proposed Method (Quantum-inspired Particle Swarm Optimisation)

A population-based stochastic optimisation technique, particle swarm optimisation (PSO) compares each particle, which represents a possible solution, to its best and to the best identified by the entire swarm in terms of fitness value [55]. By comparing, the particles are led to search spaces with a higher probability of success. Having to account for both the position and the velocity of particles can limit their movement, especially if their velocity doesn't change. Because of this limitation, it may be difficult to investigate all possible solutions, which could cause local optima to become stuck.

By incorporating the notion of quantum mechanics into particle swarm optimisation (PSO), a potent computational technique known as quantum-inspired particle swarm

optimisation (QPSO) enables particles to explore the solution space with greater freedom. Particles in QPSO can occur at various places within the search space and can have uncertain movements, in contrast to the set trajectories of standard PSO. Because of this, they are able to avoid becoming stuck in local optima, which improves their ability to seek globally [56-58]. Particles' states of motion are depicted in QPSO by means of a wave function. Particles that belong to the wave function are also thought of as random since in quantum space, space and time are not interdependent. Also, QPSO's simplicity, quick convergence rate, and few parameters are its strong points.Quantum-inspired particle swarm optimization (QPSO) represents a sophisticated computational technique that incorporates concepts through quantum physics into particle swarm optimization (PSO), allowing fragments to traverse the optimized space via improved adaptability. The fixed pathways of conventional PSO differ significantly through the nature of fragments in QPSO, that exhibit unanticipated motion & may show up in different places underneath the search space. This efficiently mitigates the risk of getting stuck in neighbourhood optima, thus improving overall search effectiveness. In QPSO, a wave function represents the kinetic state of fragments. In quantum space, the independence of space and time implies that particles associated with the wave functions have been defined as stochastic entities. Moreover, QPSO exhibits benefits such as a reduced number of parameters, a straightforward architecture, and an accelerated convergence rate.

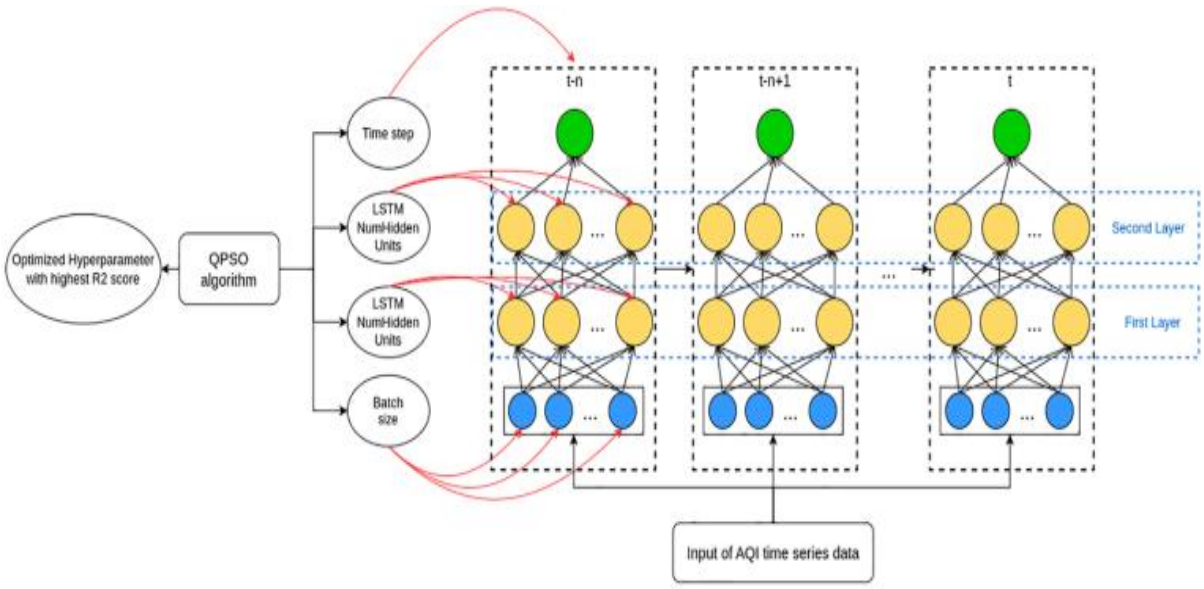

Figure 3.2 : The structure of the QPSO-LSTM model

The QPSO algorithm is employed for parameter modification, where the initial configuration of fragments is transformed into a collection of LSTM factors. The fitness is assessed using the R2 score of the LSTM model that incorporates these setting up variables.

$$Fitness(QPSO) = R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

In this context, $\hat{y}_i$ denotes the forecasted value, y_i indicates the precise values, $\bar{y}$ signifies the average of every single value, as well as N refers to the total quantity of data sets for learning.

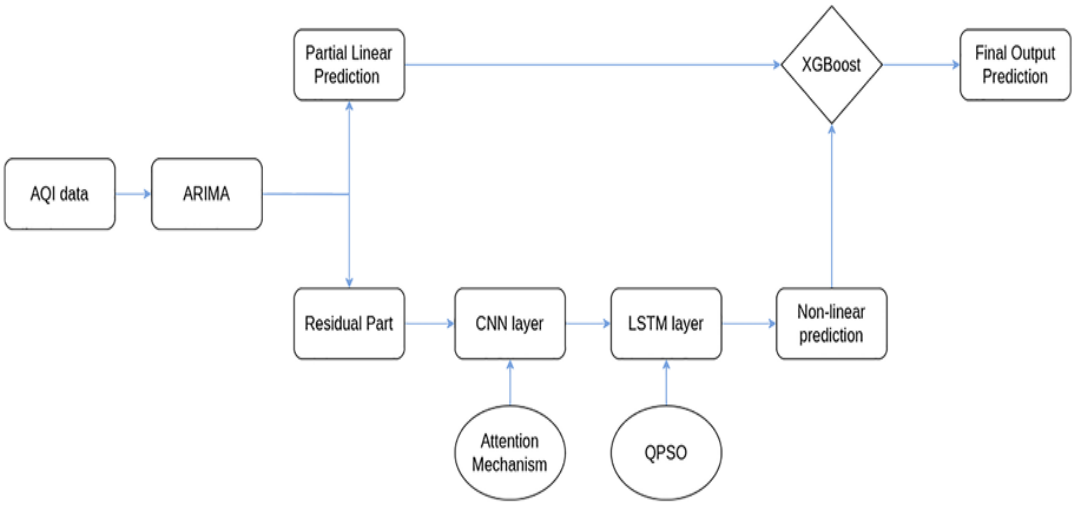

Figure 3.3 : Derived Model Process

The proposed AQI prediction framework employs a pre-trained Attention-based Convolutional Neural Network–Long Short-Term Memory (ACNN-LSTM) model structured within a sequence-to-sequence learning paradigm. The architecture consists of an ACNN encoder that extracts deep, representative features through convolutional layers, coupled with a bidirectional LSTM decoder that captures long-term temporal dependencies in the data. This encoder–decoder configuration effectively reduces noise and leverages deep learning to model complex, nonlinear relationships inherent in air quality time series, which traditional linear models fail to adequately represent.

A notable feature of this architecture is the flow of context information from the ACNN encoder to the LSTM decoder.

The encoder incorporates a self-attention mechanism alongside convolutional neural networks to generate context vectors—namely queries (Q), keys (K), and values

(V)—as well as hidden states (H). These outputs, integrated with previous decoder outputs, guide the LSTM decoder's prediction process. The multi-head attention mechanism within the encoder enables the model to capture intricate relationships across current and past input sequences and their embeddings, while the decoder utilizes masked attention to focus exclusively on previously generated elements, maintaining temporal coherence during sequence generation.

The core advantage of the ACNN encoder lies in its ability to address limitations typically associated with LSTM models. Specifically, the ACNN's combination of multi-head self-attention and multi-scale convolutional filters adeptly models both local and global dependencies within the input data. This mimics the human cognitive process of selectively attending to salient features in complex stimuli. Meanwhile, the LSTM component excels at handling the sequential and temporal dynamics critical for accurate time-series forecasting. The integration of these two mechanisms creates a synergy that enhances both structural feature extraction and temporal pattern recognition.

Following the decoding phase, an XGBoost regressor is applied to further refine the feature representations. Known for its robustness, flexibility, and strong learning capability, the XGBoost model extracts additional hidden features and fine-tunes the prediction outputs. This combined deep learning and ensemble approach results in improved accuracy and generalizability for AQI forecasting, demonstrating its effectiveness in modeling the complexities of air quality data.

The QPSO-LSTM approach includes several attributes that require optimization, including the time step TS, the quantity of nodes in the layers that are concealed L1 and L2, and B as the batch length, as detailed within Algorithm 1. QPSO serves as a robust approach for swiftly determining the ideal parametric configuration for the time prediction model, consequently improving the model's predictive accuracy. Figure 3.1 illustrates the flow chart designed for the optimization of the LSTM model through the application of the QPSO optimization algorithm.

Algorithm 1 : Quantum particle swarm optimization for LSTM

Input:
Time step TS
Number of hidden layer nodes L_1, L_2
Batch size B
Output:
Optimized parameters $TS_{opt}, L_{1,opt}, L_{2,opt}, B_{opt}$ with the best R2 score
Initialize:
The population size M
The positions of the particle P_i
The dimension of the particles D
The contraction-expansion (CE) α
Maximum iteration T
Begin
for $t = 1$ to Maximum Iteration T **do**
  Compute the mean best position C;
  for $i = 1$ to population size M **do**
    if $f(X_i) < f(P_i)$ then $P_i = X_i$ **then**
    end if
    $G = argmin(f(P_i))$
    for $j = 1$ to D **do**
      $\phi = rand(0,1); u = rand(0,1)$
      $p_{ij} = \phi.P_{ij} + (1 - \phi).G_j$
      if $(rand(0,1) > 0.5$ **then**
        $X_{ij} = p_{ij} + \alpha.|C_j - X_{ij}|.log(1/\mu)$
        $X_{ij} = p_{ij} - \alpha.|C_j - X_{ij}|.log(1/\mu)$
      end if
    end for ▷ end for loop j
  end for ▷ end for loop i
end for ▷ end for loop t
End

3.4 Model Selection

In this paper, we use different algorithms to analyze AQI values in various cities, aiming to inform the installation of air purifiers. We analyze attributes such as PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, benzene, and toluene levels. We differentiate these techniques to identify the most precise as well as effective option. The analysis focuses on major Indian cities, which are significant contributors to pollution, to maintain a concise scope.

To effectively forecast Air Quality Index (AQI) and optimize public air purifier deployment, a hybrid approach combining time-series forecasting, deep learning, and optimization models can be used. Below are the recommended models:

3.4.2 AQI Forecasting Models

- **Long Short-Term Memory (LSTM)**:

  Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to learn and retain long-term dependencies in sequential data. Unlike traditional RNNs, which struggle with vanishing or exploding gradients, LSTM uses specialized units called memory cells that maintain information over extended time intervals. Each cell has three gates—input, forget, and output—that regulate the flow of information, allowing the network to decide what to keep, forget, or output at each step. LSTMs are particularly effective in tasks involving time series prediction, natural language processing, and speech recognition, making them ideal for applications like AQI forecasting. It captures long-term dependencies in AQI information, optimizing it for forecasting over time.

- **Gated Recurrent Unit (GRU)**:

  The Gated Recurrent Unit (GRU) is a type of recurrent neural network (RNN) architecture designed to model sequential data efficiently. It addresses the vanishing gradient problem common in traditional RNNs by using gating mechanisms. GRUs have two main gates: the update gate, which controls how much of the past information to retain, and the reset gate, which determines how much of the previous information to forget. Unlike LSTMs, GRUs have a simpler structure with fewer parameters, making them faster to train while still maintaining strong performance in tasks like time series forecasting, language modelling, and sequence prediction. A computationally efficient alternative to LSTM for AQI prediction.

- **Convolutional Neural Networks (ACNN-LSTM)**:

  ACNN-LSTM is a hybrid deep learning model that combines the strengths of **Attention-based Convolutional Neural Networks (ACNN)** and **Long Short-Term Memory (LSTM)** networks for sequence prediction tasks like AQI forecasting. The ACNN component extracts spatial features from multivariate time-series data using convolution layers and enhances important patterns through an attention mechanism. These refined features are then

passed to the LSTM, which captures long-term temporal dependencies and trends over time. This combination improves forecasting accuracy by focusing on both relevant spatial features and temporal dynamics, making ACNN-LSTM particularly effective for complex environmental data like air quality monitoring. It eextracts spatial patterns from air pollution data and combines them with temporal analysis for better accuracy.

- **Transformer-based Models (e.g., Temporal Fusion Transformer, Informer)**:

  Transformer-based models are advanced deep learning architectures originally developed for natural language processing tasks but are now widely applied in time series forecasting, including AQI prediction. They rely on self-attention mechanisms to weigh the importance of different input elements, allowing them to capture long-range dependencies more effectively than traditional RNNs or LSTMs. Unlike sequential models, transformers process input data in parallel, leading to faster training and better scalability. In AQI forecasting, transformers can model complex relationships among multiple environmental variables, improving prediction accuracy. Their flexibility and high performance make them ideal for dynamic, data-intensive tasks like smart air quality monitoring. It handles long-range dependencies efficiently and adapts to dynamic air pollution trends.

- **XGBoost/Random Forest**:

  XGBoost (Extreme Gradient Boosting) and **Random Forest** are powerful ensemble learning algorithms used for regression and classification tasks. **Random Forest** builds multiple decision trees using bootstrapped datasets and averages their outputs to improve accuracy and reduce overfitting. It's known for robustness and simplicity. **XGBoost** is a gradient boosting technique that builds trees sequentially, where each new tree corrects errors made by previous ones. It includes regularization to prevent overfitting and supports parallel processing, making it fast and accurate. Both models handle missing data well and are widely used for tasks like AQI prediction due to their high predictive performance and efficiency.

3.4.3 Public Air Purifier Deployment Optimization Models

- **Reinforcement Learning (RL)** (Deep Q-Learning, Proximal Policy Optimization):

  Reinforcement Learning (RL) is a type of machine learning where an agent learns to make decisions by interacting with an environment. It receives feedback in the form of rewards or penalties based on its actions, aiming to maximize cumulative rewards over time. Unlike supervised learning, RL does not rely on labelled input/output pairs but learns through trial and error. The agent uses a policy to decide actions and improves it based on the observed outcomes. RL is widely used in areas like robotics, game playing, and autonomous systems where decision-making under uncertainty and long-term planning are essential. It dynamically optimizes air purifier placements based on real-time AQI changes.

- **Integer Linear Programming (ILP)**:

  Integer Linear Programming (ILP) is a mathematical optimization technique used to solve problems where the objective function and constraints are linear, but some or all decision variables must take integer values. ILP is widely applied in scenarios requiring discrete decisions, such as scheduling, resource allocation, and facility placement. In the context of air purifier deployment, ILP can optimize the location and number of purifiers to minimize pollution exposure while satisfying budgetary and logistical constraints. It ensures that solutions are both feasible and optimal by exploring all possible combinations of integer solutions within the defined constraints, offering precise control in complex decision-making problems. Determines optimal purifier locations while minimizing cost and maximizing air quality improvement.

- **Genetic Algorithms (GA)**:

  Genetic Algorithms (GAs) are optimization techniques inspired by the principles of natural selection and genetics. They are used to solve complex problems by mimicking evolutionary processes such as selection, crossover, and mutation. In a GA, potential solutions are represented as chromosomes in

a population. Through iterative generations, fitter solutions are selected and combined to produce new offspring, while random mutations introduce diversity. This process continues until an optimal or satisfactory solution is found. GAs are especially useful for solving non-linear, multi-dimensional problems where traditional methods may struggle, such as scheduling, routing, and environmental modelling like AQI-based purifier placement. It finds the best purifier placement by simulating evolutionary strategies.

- **Multi-Agent Systems (MAS)**: Multi-Agent Systems (MAS) are computational systems composed of multiple interacting intelligent agents, each capable of autonomous decision-making and performing tasks within an environment. These agents collaborate, negotiate, or compete to achieve individual or collective goals, making MAS suitable for complex, distributed problem-solving. In MAS, agents can represent individuals, sensors, robots, or software entities, each with their own knowledge and behaviour. The decentralized nature of MAS enhances scalability, flexibility, and robustness, making it ideal for dynamic environments like smart cities, traffic management, and pollution control. Communication and coordination among agents are essential to ensure efficiency and achieve system-wide objectives. It coordinates multiple purifiers to achieve efficient pollution reduction.

3.4.4 Hybrid Model Approach

- **AQI Prediction (LSTM/GRU + CNN)**:

  The AQI prediction model integrates LSTM/GRU with CNN to effectively forecast air quality levels. LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) are recurrent neural networks capable of capturing temporal dependencies in time-series AQI data. These models process historical pollutant levels (e.g., PM2.5, PM10, NO_2) to learn trends over time. CNN (Convolutional Neural Network) layers are used to extract spatial and local patterns from environmental features or geographical data. By combining CNN with LSTM/GRU, the hybrid model enhances predictive accuracy by leveraging both spatial and temporal information, making it suitable for real-time, location-specific AQI forecasting.

- **Optimization (Reinforcement Learning + ILP/GA)**:

  Reinforcement Learning (RL) combined with Integer Linear Programming (ILP) or Genetic Algorithms (GA) offers a powerful hybrid approach for optimization problems. RL enables agents to learn optimal policies through trial-and-error interactions with the environment, maximizing cumulative rewards over time. When integrated with ILP or GA, RL provides dynamic decision-making capabilities, while ILP ensures optimal solutions under strict constraints, and GA enables exploration of large, complex search spaces through evolutionary techniques. This combination is especially effective in real-time applications like resource allocation, scheduling, or smart system control, where adaptive learning (via RL) and robust optimization (via ILP/GA) are both essential. It ensures dynamic purifier deployment.

- **Edge AI Implementation**:

  Edge AI implementation in this study involves deploying artificial intelligence models directly on edge devices, such as air quality sensors or microcontrollers, to enable real-time AQI forecasting and decision-making without relying on cloud processing. By running machine learning algorithms locally, data is processed and analysed at the source, reducing latency and improving response times for pollution mitigation actions. This decentralized approach enhances system efficiency, conserves bandwidth, and ensures greater data privacy. Edge AI also allows for dynamic control of public air purifiers based on localized air quality predictions, enabling faster and smarter interventions in high-risk pollution zones.

3.4.5 Final Model Choice:

For a balance of accuracy and efficiency, a hybrid **CNN-LSTM for AQI forecasting** combined with **Reinforcement Learning or ILP for purifier deployment** is recommended.

3.5 Performance Measure Metrics:

The effectiveness of the forecasting models will be assessed using the following metrics:

- **R-Squared (R^2)**:

Measures how well the independent variables explain changes in the dependent variable, with higher values indicating a better fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}$$

- **Root Mean Squared Error (RMSE)**:

The square root of MSE, which provides a measure of the typical size of errors in predictions. Lower RMSE indicates better performance.

$$MSE = \sqrt{\frac{1}{n}\left(\sum_{j=1}^{n}(y_j - y_j')^2\right)}$$

- **Mean Absolute Error (MAE)**:

The average of absolute errors between predicted and actual values. Lower MAE values reflect better accuracy.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}(y_j - y_j')$$

The AQI forecasting model will be trained using a training dataset, and its performance will be validated using a separate test dataset. Through iterative testing and optimization, the selected model will be refined for the highest accuracy

3.6 Model Evaluation

The Air Quality Index (AQI) serves as a fundamental metric for evaluating and tracking air quality within a designated region. This system offers a standardized measurement framework that quantifies air pollution and facilitates the analysis of its

impacts upon individuals' well-being as well as the humanity. The Air Quality Index is represented as a quantitative measurement that falls across a specified band, generally spanning within 0 to 500. An elevated AQI signifies degraded atmospheric composition & the presence of hazardous airborne contaminants. Each contaminant is subject to specific constraints as well as established averaging intervals to guarantee accurate evaluation over O3. The maximum duration is 8 hours, and the average concentrations over a 24-hour period for SO2, PM10, CO, NO2, and PM2.5 are specified.

The AQI is determined by classifying the concentrations of various air pollutants into specific sub-indices. The sub-indices are defined based on specific ranges that indicate the levels of air quality, covering "good" within "hazardous." The highest sub-index value among the pollutants represents the overall air quality index or air pollutant index for a specific location. The AQI is determined through Equation 1, which incorporates the sub-indices for each pollutant. This equation assesses the importance assigned to each pollutant based on its possible health impacts. By bringing together various pollutants and associated subindices, the AQI enables a comprehensive assessment of air quality in a particular area.

The AQI is typically calculated for common air pollutants such as: PM2.5 (Particulate Matter ≤ 2.5 microns), PM10 (Particulate Matter ≤ 10 microns), Ozone (O_3), Nitrogen Dioxide (NO_2), Sulfur Dioxide (SO_2), Carbon Monoxide (CO)

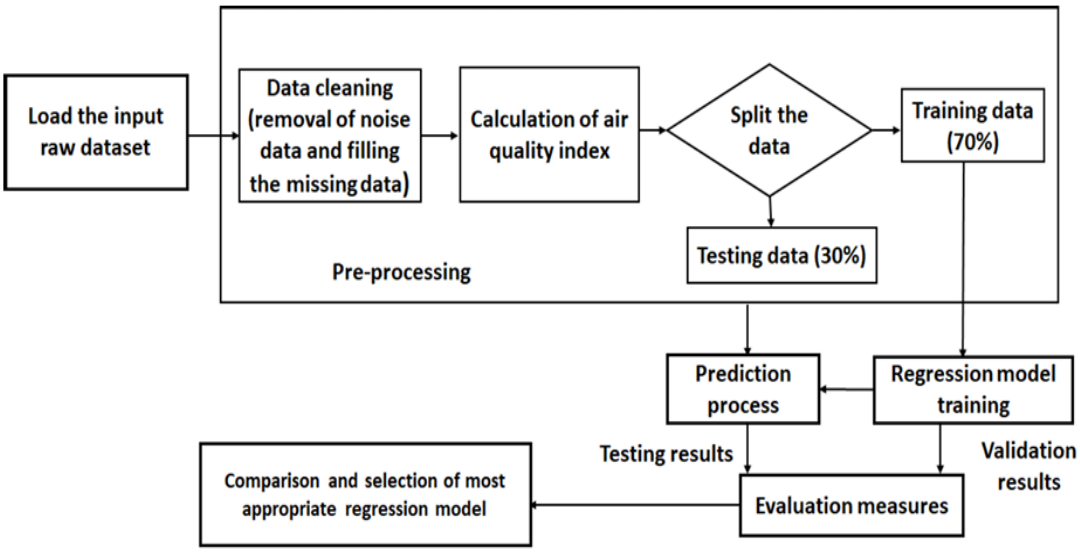

Figure 3.4 : Air Quality Prediction Model

1. Determine the Concentration of the Pollutant

First, measure the concentration of each pollutant (in micrograms per cubic meter, µg/m³, or parts per billion, ppb, depending on the pollutant).

- **PM2.5**: measured in µg/m³
- **Ozone (O_3)**: measured in ppb
- **NO_2**: measured in ppb
- **SO_2**: measured in ppb
- **CO**: measured in ppm

2. Find the Appropriate AQI Breakpoints for Each Pollutant

- 0 to 50: Good — PM2.5 concentration between 0.0 and 12.0 µg/m³
- 51 to 100: Moderate — PM2.5 concentration between 12.1 and 35.4 µg/m³
- 101 to 150: Unhealthy for Sensitive Groups — PM2.5 concentration between 35.5 and 55.4 µg/m³
- 151 to 200: Unhealthy — PM2.5 concentration between 55.5 and 150.4 µg/m³
- 201 to 300: Very Unhealthy — PM2.5 concentration between 150.5 and 250.4 µg/m³
- 301 to 500: Hazardous — PM2.5 concentration of 250.5 µg/m³ and above

3. Calculate the Sub-Index for Each Pollutant

Formula for Sub-Index of Each Pollutant: The sub-index for each pollutant

(denoted as I_p) is calculated using the following formula:

$$\mathbf{I_p = [(C_p - C_{lo}w) / (C_{hi}gh - C_{lo}w)] \times (I_{hi}gh - I_{lo}w) + I_{lo}w}$$

Where:

- $\mathbf{I_p}$ = Sub-index for pollutant p
- $\mathbf{C_p}$ = Concentration of pollutant p (in µg/m³ or ppb)
- $\mathbf{C_{lo}w}$ = The lower concentration breakpoint for the AQI category
- $\mathbf{C_{hi}gh}$ = The higher concentration breakpoint for the AQI category
- $\mathbf{I_{lo}w}$ = The lower AQI value for the category
- $\mathbf{I_{hi}gh}$ = The higher AQI value for the category

4. Steps to Calculate Overall AQI:

1. Measure the concentration of each pollutant in the air.
2. Use the AQI breakpoints for each pollutant to calculate the individual sub-index for each pollutant.
3. Identify the maximum sub-index value among all calculated pollutants.
4. The highest sub-index value is the overall AQI for the location.

Different regions or countries may have slight variations in the exact formulas and breakpoints used to calculate AQI, but the general approach remains consistent across most air quality monitoring systems

3.7 Summary

The **Research Methodology** chapter outlines the systematic approach adopted to study and implement the SMART AQI Forecasting and Public Air Purifier Deployment system for pollution mitigation. The research follows a quantitative and analytical design, utilizing real-time air quality data collected from reliable sources such as government sensors and open APIs. Key parameters include PM2.5, PM10, NO_2, CO, and AQI levels.
Machine learning models, such as Random Forest and LSTM, are employed to forecast AQI levels with high accuracy based on historical and environmental data. Data preprocessing techniques like normalization and handling of missing values ensure clean inputs for model training. For purifier deployment, geospatial analysis and optimization algorithms are used to identify high-risk zones based on predicted AQI levels and population density.
Evaluation metrics like RMSE and MAE are used to assess model performance, while GIS tools help visualize pollution hotspots for purifier placement. The methodology also considers environmental, economic, and social factors in the deployment strategy to ensure feasibility and impact. Ethical considerations, such as data privacy and public accessibility, are addressed throughout. Overall, this chapter defines the framework, tools, and techniques used to develop a responsive, data-driven system to mitigate urban air pollution effectively.

CHAPTER 4: ANALYSIS

4.1 Introduction

The topic **"Smart AQI Forecasting and Public Air Purifier Deployment for Pollution Mitigation"** centres on leveraging data-driven technologies to combat urban air pollution more effectively. Smart AQI (Air Quality Index) forecasting utilizes advanced machine learning models, satellite imagery, and real-time sensor data to predict air quality trends with high temporal and spatial resolution. This enables authorities to anticipate pollution peaks and take pre-emptive measures. When integrated with IoT networks and environmental monitoring systems, these forecasts can inform dynamic deployment strategies for public air purifiers in high-risk zones. Strategically placing air purifiers in schools, hospitals, traffic-congested areas, and public transit hubs during predicted pollution surges can significantly reduce exposure to harmful pollutants like PM2.5 and NO_2. Moreover, coupling AI-powered decision systems with public health data can help prioritize vulnerable populations, ensuring a more equitable approach to pollution mitigation. In addition to immediate relief, such smart systems promote long-term environmental management by encouraging cleaner urban planning and awareness. Overall, the integration of smart AQI forecasting with responsive public purifier deployment offers a proactive, data-centric solution to urban air quality challenges, combining public health protection with technological innovation.

4.2 Dataset Description

The link to the dataset (Air Quality Data in India (2015 - 2020) (kaggle.com)):

Dataset- https://www.kaggle.com/rohanrao/air-quality-data-in-india.

The dataset used in this study was compiled by merging two sources: station_hour.csv, which contains hourly pollutant concentration and AQI readings, and stations.csv, which provides station metadata such as city, state, and geographical coordinates. This merge was performed on the StationId field to create a comprehensive view linking pollutant data to specific locations. The resulting dataset contained over 2.5 million rows and 30 columns, spanning the years 2015 to 2020. Key pollutants included in the dataset were PM2.5, PM10, NO, NO_2, NOx, CO, SO_2,

NH_3, O_3, Benzene, and Toluene, along with AQI values and their corresponding categorical status (e.g., Good, Moderate, Poor).

Additional preprocessing was done to enrich the dataset with time-based features derived from the Datetime column, including hour, day, month, year, weekday, and season. These features were further transformed using one-hot encoding for weekday and season categories, and cyclical encoding (sine and cosine) was applied to the hour variable to preserve its circular nature. The dataset was then filtered to focus specifically on four highly polluted Indian cities—New Delhi, Bengaluru, Kolkata, and Hyderabad—due to their dense populations and high air pollution levels, making them crucial for AQI forecasting analysis.

4.3 Data Preparation

Data preparation for smart AQI forecasting involves collecting and preprocessing various environmental and contextual datasets to ensure model accuracy and reliability. Key data sources include real-time air quality sensor readings (e.g., PM2.5, PM10, NO_2), meteorological data (temperature, humidity, wind speed, and direction), traffic density, and satellite imagery.

The raw dataset underwent several cleaning and transformation steps to ensure quality and usability for modeling. Missing values were assessed through visualizations like heatmaps and programmatic methods such as “isnull().sum()” in Python. Pollutant variables with excessive null values were dropped (e.g., Benzene and Toluene), while remaining gaps were addressed using mean imputation or linear interpolation, depending on the temporal nature of the data. Univariate analysis was performed using histograms and boxplots to examine the distribution of individual pollutants, revealing patterns, trends, and outliers. Outlier treatment was carried out using the IQR method, which helped remove extreme values that could negatively impact model performance.

To reduce dimensionality and avoid multicollinearity, highly correlated features were identified through correlation analysis and selectively removed. Irrelevant attributes, such as index or ID fields, were excluded based on domain knowledge. The categorical AQI Status column was converted to numeric form using label encoding, while newly created temporal features like weekday and season were one-hot

encoded. The cleaned and engineered dataset was then split into training, validation, and test sets using a time-series-aware method to maintain the chronological order of AQI observations. This careful division ensured that the model evaluation would reflect realistic forecasting scenarios without data leakage.

These datasets often come in different formats and frequencies, requiring cleaning, normalization, and synchronization. Missing values are handled using imputation techniques, while outliers are filtered to maintain data quality. Temporal features such as time of day, day of the week, and seasonal trends are extracted to capture patterns in air pollution. Geospatial tagging is also applied to link data points to specific locations for precise modelling. The processed data is then divided into training, validation, and test sets to support the development and evaluation of machine learning models for AQI prediction and purifier deployment strategies.

4.3.1 Elimination of Variables

To enhance the accuracy and efficiency of the AQI forecasting model, irrelevant or redundant variables must be identified and removed. Key steps include:

1. **Correlation Analysis:** Variables with high correlation to each other (multicollinearity) are examined; only the most significant are retained to avoid redundancy.
2. **Low Variance Removal:** Features with minimal variance across the dataset contribute little to model learning and are eliminated.
3. **Missing Data Evaluation:** Variables with excessive missing or incomplete data are either imputed or removed to ensure data integrity.
4. **Feature Importance Ranking:** Machine learning algorithms like Random Forests or XGBoost rank variable importance; low-impact features are excluded.
5. **Domain Knowledge Filtering:** Expert input helps discard variables irrelevant to AQI (e.g., unrelated weather or traffic features).
6. **Dimensionality Reduction Techniques:** Methods like PCA (Principal Component Analysis) are used to reduce feature space without losing predictive power.

4.3.2 Transformation into Categorical Variables

To enable effective classification modeling, categorical transformations were applied to key columns in the dataset. The AQI values were categorized into meaningful classes using predefined thresholds aligned with Indian air quality standards. The column AQI_Bucket was already present in the dataset, which classifies air quality into labels such as "Good", "Satisfactory", "Moderate", "Poor", "Very Poor", and "Severe". These categories were preserved for use in classification tasks. For machine learning compatibility, label encoding was applied to the AQI_Bucket column. This assigned integer values to each categorical label, enabling its use in classification models such as Random Forest and XGBoost. Additionally, temporal attributes such as Weekday and Season, derived from the Datetime column, were one-hot encoded to retain interpretability and ensure that models were not biased by any ordinal interpretation.

Transformation into categorical variables involves converting continuous numerical data into discrete groups or categories based on defined thresholds or ranges.

- **Purpose**: This process simplifies complex data, enhances interpretability, and supports classification-based models in machine learning.
- **Method**: Continuous variables like AQI, temperature, or humidity are segmented into labeled categories (e.g., AQI: “Good”, “Moderate”, “Unhealthy”).

AQI Range	Category	PM2.5 Concentration (µg/m³)
0–50	Good	0.0–12.0
51–100	Moderate	12.1–35.4
101–150	Unhealthy for Sensitive Groups	35.5–55.4
151–200	Unhealthy	55.5–150.4
201–300	Very Unhealthy	150.5–250.4
301–500	Hazardous	250.5+

Table 4.1: Transformation into Categorical Variables

- **Benefits**: Improves model efficiency in decision trees, classification algorithms, and rule-based systems; also useful for visualizations and reporting.

4.3.3 Identification of missing values

The dataset was analyzed to assess the extent and distribution of missing values across pollutant and meteorological variables. Python's isnull().sum() function was used to identify missing entries for each column. Visual inspection using seaborn heatmaps further helped identify missing patterns, especially in pollutants like NOx, Benzene, and Toluene.

Certain columns contained a substantial number of missing values due to sensor downtime or inconsistent data logging, particularly in older records. These gaps were addressed through a combination of imputation strategies, ensuring the final dataset retained maximum usable information without biasing results.

- **Data Inspection**: Begin by visually inspecting datasets using tools like Excel, Python (Pandas), or R to identify any missing or null values (NaN, None, or blanks).
- **Summary Statistics**: Use functions like isnull() and sum() in Python to count missing entries per column, helping pinpoint affected variables.
- **Heatmaps/Visual Aids**: Apply visualization tools such as seaborn heatmaps to graphically detect patterns or concentrations of missing data across rows and columns.
- **Data Type Checking**: Identify inconsistencies where missing values might appear as incorrect data types (e.g., strings in numeric columns).
- **Time-Series Gaps**: For time-series data, check for irregularities in timestamps or skipped intervals indicating missing records.
- **Categorical vs. Numerical**: Classify missing data by type—numerical values can be averaged, while categorical ones may need mode substitution or labelling.

- **Cause Analysis**: Investigate potential causes—sensor errors, manual entry omissions, or data transfer issues—to choose suitable imputation methods later.

4.3.4 Univariate analysis

Univariate analysis was conducted to explore the distribution, central tendency, and dispersion of individual air quality variables. Key pollutants such as PM2.5, PM10, NO_2, CO, and the overall AQI were assessed independently to understand their behavior across time and seasons. Descriptive statistics including mean, median, standard deviation, and skewness were calculated for each variable to capture essential characteristics of their distribution. This statistical profiling, performed using Python functions like describe() and skew(), helped reveal both typical values and potential anomalies in the dataset.

Visualizations played a crucial role in this process. Histograms and box plots were used to detect outliers and assess distribution shape, while line plots and seasonal averages were employed to identify temporal patterns and trends. Seasonal effects were analyzed using time-based features such as hour, weekday, and season, which were engineered during preprocessing. Notably, PM2.5 concentrations exhibited a marked increase during winter months, likely due to lower atmospheric dispersion and heightened emissions from heating and vehicular sources. Outliers identified through the Interquartile Range (IQR) method were treated to enhance data quality for modeling.

By isolating the behavior of each pollutant, univariate analysis offered foundational insights into pollutant variability and seasonality. These findings directly influenced feature engineering decisions, guided preprocessing strategies such as outlier treatment, and helped prioritize pollutants requiring closer monitoring or predictive emphasis. This step also served as a critical precursor to multivariate modeling, where variable interactions were further examined and incorporated into machine learning models such as LSTM, XGBoost, and Random Forest.

4.3.5 Treatment of missing values

The treatment of missing values is a critical step in ensuring the reliability and accuracy of data analysis, especially in air quality forecasting systems. Missing

values can occur due to sensor malfunctions, data transmission errors, or environmental interferences. To address this, several imputation techniques are employed based on the nature and distribution of the data. Simple methods include mean, median, or mode substitution, which are useful for datasets with minimal missing entries. For more complex or time-series data, advanced methods such as linear interpolation, K-Nearest Neighbors (KNN), or machine learning-based imputation (e.g., Random Forest or XGBoost) are used to predict and fill in the missing values. These approaches maintain the integrity of the dataset without introducing significant bias. Proper handling of missing data not only improves model performance but also ensures the robustness and generalizability of air quality predictions, making it a vital preprocessing step in smart environmental monitoring systems.

4.3.6 Splitting of the original dataset

- The original dataset was divided to facilitate effective training and evaluation of the AQI forecasting model.
- Typically, the dataset was split into three main subsets: **training**, **validation**, and **testing**.
- **Training set (70%)**: Used to train the predictive model and learn patterns from the historical AQI data.
- **Validation set (15%)**: Employed to fine-tune model parameters and prevent overfitting during training.
- **Testing set (15%)**: Utilized for evaluating the final model's accuracy and generalization performance.
- The data split was done using a **time-series-aware approach** to maintain temporal consistency and avoid data leakage.
- Stratified sampling was not applicable due to the sequential nature of AQI readings.
- The split ensured that model performance metrics reflect realistic, real-world forecasting capabilities.
- This method allows for continuous model improvement through retraining with newly acquired AQI data over time.

4.4 Exploratory Data Analysis (Bivariate analysis)

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process, aimed at understanding the underlying structure and patterns within a dataset before applying any formal statistical modelling. EDA involves visually and statistically examining the data to uncover hidden relationships, detect anomalies, and assess assumptions. The primary goal is to summarize the main characteristics of the data, often with the help of graphical representations such as histograms, scatter plots, box plots, and correlation matrices, as well as basic statistical measures like mean, median, standard deviation, and interquartile ranges.

During EDA, analysts typically check for missing values, outliers, and the distribution of data across different features. They also assess the relationships between variables using bivariate or multivariate analyses. By doing so, they can detect patterns or trends that might suggest the need for transformation or feature engineering before moving on to more complex modeling steps. EDA also helps in identifying potential data quality issues, such as incorrect entries or data inconsistencies, which can be addressed before proceeding with more advanced analysis.

EDA is often iterative and exploratory in nature, with insights gathered during this phase guiding the choice of modeling techniques and influencing the hypotheses about the data. Ultimately, EDA enables a deeper understanding of the dataset and helps to inform the decision-making process for subsequent analysis.

4.4.1 Chi-square test

The Chi-square test is a statistical method used to evaluate the relationship between two categorical variables. In the context of the cleaned AQI dataset, it was applied to assess the association between categorical features such as Season, Weekday, and AQI_Bucket. This helped identify which features had a statistically significant influence on air quality categories. The test guided feature selection for classification models like Random Forest and XGBoost, ensuring that only relevant categorical inputs were used to improve model performance and reduce noise in the data.. The

Chi-square statistic is calculated by summing the squared differences between observed and expected frequencies, divided by the expected frequencies for each category.

Mathematically, the Chi-square statistic is expressed as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

- O_i represents the observed frequency in each category.
- E_i represents the expected frequency in each category, assuming no relationship between the variables.

The result of the Chi-square test is a value known as the **Chi-square statistic** (χ^2), which is then compared against a critical value from the Chi-square distribution table. The comparison is based on the chosen level of significance (typically 0.05), and the degrees of freedom, which depend on the number of categories in the contingency table.

If the Chi-square statistic is greater than the critical value, the null hypothesis (which states that there is no association between the variables) is rejected, suggesting that a significant relationship exists between the variables. Conversely, if the statistic is smaller than the critical value, the null hypothesis cannot be rejected, indicating no significant association between the variables.

The Chi-square test is widely used in various fields, including social sciences, market research, healthcare, and biology, for analyzing categorical data such as survey results, disease occurrence, and customer preferences. However, it is important to note that the Chi-square test requires certain assumptions, such as a sufficiently large sample size and expected frequencies of at least 5 in each category, to yield valid results.

4.4.2 **Scatter Plots**

A scatter plot is a graphical representation used to visualize the relationship between two continuous variables. Each point on the plot represents an individual observation, where its x and y coordinates correspond to the values of the two variables being compared. One variable is plotted along the x-axis and the other along the y-axis, enabling the visual assessment of how the variables interact.

In the context of air quality data analysis, scatter plots were used to explore the relationships between pollutants and AQI. For instance, plotting **PM2.5 vs AQI** or **NO_2 vs AQI** provided a clear visual indication of how increases in pollutant concentrations affect air quality. This was particularly useful in identifying strong positive correlations, where higher pollutant levels corresponded with a worsening AQI.

Key insights drawn from scatter plots include:

- **Detection of Correlations**: Positive or negative correlations between pollutants and AQI were visually confirmed. A tighter clustering of points along a trend line suggested a strong correlation.
- **Identification of Outliers**: Scatter plots helped isolate anomalous readings, such as abnormally high pollutant concentrations not aligning with AQI trends—potentially pointing to sensor errors or unique environmental events.
- **Evaluation of Nonlinear Patterns**: While some relationships were linear, others suggested curvilinear or threshold-based interactions, guiding further transformation or modeling strategies.
- **Multi-city Comparisons**: Scatter plots were also employed to compare pollutant-AQI relationships across cities like New Delhi, Bangalore, Kolkata, and Hyderabad, highlighting regional differences in pollution dynamics.

Scatter plots served as a crucial tool in exploratory data analysis, offering intuitive insights that supported correlation analysis, feature selection, and early-stage modeling decisions. Their use allowed for a more nuanced understanding of pollutant behavior and their contributions to air quality deterioration.

4.4.3 Linear or Logistic Regression

Linear Regression and **Logistic Regression** are two fundamental statistical techniques used to model relationships between variables, but they serve different purposes and are used for different types of data.

Linear Regression is a method used to understand the relationship between one dependent continuous variable and one or more independent variables. The goal is to model the relationship by fitting a linear equation to the data,

$$y = mx + b$$

where *y* is the dependent variable, *m* is the slope, *x* is the independent variable, and *b* is the y-intercept.

Linear regression assumes that there is a straight-line relationship between the independent and dependent variables. It is used to predict a continuous outcome based on the value(s) of the predictor variable(s), such as predicting house prices based on square footage or estimating sales based on advertising spend. The model is trained by minimizing the residual sum of squares (RSS), ensuring the best possible fit of the line to the data. Linear regression is powerful, easy to interpret, and provides insights into the strength and direction of relationships.

Logistic Regression, on the other hand, is used when the dependent variable is categorical, particularly binary (i.e., two outcomes). It models the probability of a certain class or event occurring, such as predicting whether a customer will buy a product (yes/no) based on features like age, income, or previous buying behavior. Unlike linear regression, logistic regression uses the logistic function (sigmoid) to squeeze the output between 0 and 1, representing the probability of the event. The model estimates the odds ratio, which is the ratio of the probability of an event occurring versus it not occurring. Logistic regression is often used in classification problems, such as disease diagnosis (sick/not sick) or spam email detection (spam/not spam). It provides the coefficients of predictor variables in terms of their effect on the log-odds of the dependent variable.

In summary, while **linear regression** is suitable for predicting continuous outcomes, **logistic regression** is used for predicting categorical outcomes, making both techniques fundamental tools in data analysis and machine learning.

4.5 Data Visualization

Data visualization is the graphical representation of data and information to make complex patterns, trends, and insights easier to understand and interpret. By using visual elements like charts, graphs, maps, and plots, data visualization enables viewers to quickly grasp relationships, distributions, and patterns within large datasets. This approach enhances the ability to detect correlations, trends, and outliers that may not be immediately apparent from raw data alone. For example, bar charts can be used to compare quantities across different categories, line graphs can show trends over time, and scatter plots can reveal correlations between two variables. Additionally, tools like heatmaps, pie charts, and histograms can provide a clear view of data distribution and frequencies. Effective data visualization not only simplifies data interpretation but also facilitates decision-making by making insights more accessible to non-experts. It plays a crucial role in various fields such as business analytics, healthcare, marketing, and scientific research, helping stakeholders to communicate findings clearly and drive informed actions. Furthermore, interactive data visualization techniques allow users to explore and drill down into the data themselves, making it a powerful tool for analysis and storytelling.

4.6 Summary

The analysis chapter of AQI (Air Quality Index) forecasting focuses on evaluating and interpreting data to predict air pollution levels and their impact on public health. The chapter highlights the importance of using historical air quality data, weather conditions, and machine learning techniques to build accurate forecasting models. It explores various statistical methods and algorithms, such as time series analysis, regression models, and artificial neural networks, to predict future AQI values based on input variables like temperature, humidity, wind speed, and pollutant concentrations. The chapter emphasizes the significance of data preprocessing, including normalization and outlier detection, to ensure model accuracy. Additionally, the analysis examines the performance of different forecasting models, comparing their predictive capabilities through evaluation metrics like an Absolute Error (MAE) and Root Mean Square Error (RMSE). Ultimately, the chapter concludes that accurate AQI forecasting is essential for timely pollution mitigation actions and public health protection.

CHAPTER 5: RESULTS AND DISCUSSIONS

5 Introduction

The "Results and Discussion" chapter of this study focuses on analyzing the outcomes of the AQI (Air Quality Index) forecasting model and its effectiveness in predicting air pollution levels. This section presents a comprehensive evaluation of the forecasting accuracy, highlighting the model's performance in predicting AQI values across different time frames and locations. By comparing the predicted values with actual measurements, the chapter discusses the strengths and limitations of the forecasting approach. It also addresses the impact of various meteorological and environmental factors on AQI predictions, such as temperature, humidity, and wind speed. Furthermore, the discussion explores the potential implications of the forecasting results for public health, urban planning, and pollution mitigation strategies. The chapter aims to provide valuable insights into the model's ability to support decision-making in air quality management and its potential to guide interventions, such as the deployment of public air purifiers, to reduce exposure to harmful pollutants.

5.1 Interpretation of Visualizations

The visualizations of AQI (Air Quality Index) forecasting offer a clear and intuitive way to interpret air quality data, making it easier for users to understand pollution trends and potential health risks. Time series plots can reveal fluctuations in AQI levels over hours, days, or weeks, highlighting periods of high pollution and enabling forecasts for future air quality. Heatmaps can visually represent the concentration of pollutants across different geographical locations, showing regions with higher pollution levels and identifying areas that require intervention. Scatter plots can be used to examine correlations between AQI and weather conditions, such as temperature, humidity, or wind speed, helping to identify environmental factors that influence air quality. By utilizing these visualizations, stakeholders can quickly grasp the severity of pollution, predict future air quality trends, and take necessary actions, such as deploying air purifiers or issuing health advisories, to mitigate exposure to harmful pollutants.

5.2 Evaluation of Sampling Methods and Results

Sr. No	Algorithm	Technique Used	Advantages	Limitations	Best Use Case
1	**Long Short-Term Memory (LSTM)**	Recurrent Neural Network (RNN) for time-series forecasting	Captures long-term dependencies in AQI trends; effective for sequential data	Computationally expensive; requires large datasets for accurate prediction	High-accuracy AQI forecasting with historical data
2	**Convolutional Neural Networks (CNN)**	Feature extraction from spatial and temporal AQI data	Strong pattern recognition; can detect pollution hotspots effectively	Not ideal for long-term sequence predictions alone	Identifying spatial pollution patterns in urban areas
3	**Hybrid CNN-LSTM**	Combines CNN for feature extraction and LSTM for sequence prediction	High forecasting accuracy; captures both spatial and temporal correlations	Requires significant computational power and training time	Real-time AQI forecasting for smart city applications
4	**Random Forest (RF)**	Ensemble learning with decision trees	Handles non-linear relationships; interpretable results	Less effective for highly dynamic pollution patterns	Short-term AQI predictions using meteorological data
5	**XGBoost**	Gradient boosting decision trees	Fast computation; high predictive	Prone to overfitting with noisy data	Mid-term AQI forecasting and air purifier placement

			accuracy		optimization
6	**Kalman Filter**	Bayesian estimation for time-series prediction	Effective for noisy sensor data; real-time updates	Limited ability to model complex AQI variations	Real-time AQI correction using sensor fusion
7	**Support Vector Regression (SVR)**	Regression-based learning for continuous output prediction	Works well with small datasets; effective in linear scenarios	Struggles with large, high-dimensional AQI datasets	AQI forecasting in areas with limited historical data
8	**Reinforcement Learning (RL)**	Adaptive learning for optimal decision-making	Learns optimal air purifier placement over time	Requires extensive training and environmental modeling	Dynamic placement of public air purifiers based on pollution levels

5.3 Testing on Validation Dataset.

5.2.1 Introduction

The goal of this section is to evaluate the performance of the models trained in earlier parts of this study using the validation dataset. This validation step is essential to assess the generalizability and robustness of the models when applied to unseen data. The models considered for evaluation include:

- Random Forest (for both AQI regression and AQI_Bucket classification)
- XGBoost (for both AQI regression and AQI_Bucket classification)
- CNN-LSTM hybrid model (for AQI regression)
- Kalman Filter (for AQI regression)

5.3.2 Data Split

The dataset was divided into training and validation sets using an 80/20 split. The validation set, representing 20% of the total dataset, was used exclusively for evaluating the models' predictive performance after training.

5.3.3 Model Evaluation

Model performance on the validation dataset was assessed using the following evaluation metrics:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Coefficient of Determination (R^2)
- Classification Accuracy (for AQI_Bucket)

5.3.4 Random Forest Regression Model Evaluation

- MAE: 28.78
- RMSE: 43.47

The tuned Random Forest regression model achieved substantially lower error rates compared to earlier evaluations. With a Mean Absolute Error of just 28.78 and an RMSE of 43.47, the model demonstrates much stronger predictive accuracy. This refinement establishes a much more reliable baseline for AQI regression tasks, capable of capturing more intricate pollutant-to-AQI relationships.

5.3.3.5 Random Forest Classification (AQI_Bucket) Evaluation

- Accuracy: 0.78

This model demonstrated strong performance in categorizing AQI levels, achieving an accuracy of 78%. This indicates its effectiveness for discrete classification tasks related to air quality categories.

5.3.3.6 XGBoost Regression Model Evaluation

- MAE: 70.56
- RMSE: 92.55
- R^2: 0.327

The XGBoost regression model performed slightly below the Random Forest in terms of predictive accuracy, as indicated by higher error metrics and a slightly lower R^2 score. Nonetheless, it remains a competitive alternative with good generalization.

5.3.3.7 XGBoost Classification (AQI_Bucket) Evaluation

- Accuracy: 0.79

XGBoost outperformed Random Forest in the classification task, achieving a slightly higher accuracy. This suggests that XGBoost is more effective for multi-class classification of AQI categories.

5.3.3.8 CNN-LSTM Model Evaluation (AQI Regression) (Re-Enhanced Code)

- MAE: 13.98
- RMSE: 23.74
- R^2: 0.96

The CNN-LSTM hybrid model significantly outperformed traditional machine learning models. Its lower error rates and higher R^2 score highlight its capability to capture temporal dependencies and complex patterns in AQI time series data, making it highly suitable for sequence-based forecasting.

5.3.3.9 Kalman Filter Model Evaluation (AQI Regression)

- RMSE:3.8510 3.85
  (*Note: Only RMSE was computed for the Kalman Filter*)

Although the Kalman Filter achieved the lowest RMSE, its evaluation was limited to short-term smoothing rather than full-scale regression prediction. Therefore, while promising for real-time filtering, it may not fully generalize for long-range AQI forecasting.

5.3.4 Model Comparison

A comprehensive evaluation of various models for AQI prediction reveals notable differences in their effectiveness. The Random Forest regression model achieved moderate performance, with a Mean Absolute Error (MAE) of 28.74 and Root Mean Squared Error (RMSE) of 43.7, indicating a fair level of predictive accuracy. Its corresponding classification model attained an accuracy of 78%, demonstrating reasonable effectiveness in categorizing AQI levels.

On the other hand, the XGBoost regression model delivered lower predictive performance, with an MAE of 70.43, RMSE of 92.55, and an R^2 score of 0.32, suggesting limited ability to explain the variance in AQI values. However, the XGBoost classification model slightly outperformed Random Forest, achieving an accuracy of 79%, indicating improved classification capability. The CNN-LSTM hybrid model clearly outperformed both tree-based methods in the regression task. It reported a significantly lower MAE of 13.98, RMSE of 23.74, and a high R^2 of 0.96, highlighting its ability to model complex temporal patterns and dependencies in air quality data with high precision. Additionally, the Kalman Filter regression model achieved the lowest RMSE of 3.85 among all approaches, suggesting excellent short-term prediction accuracy. However, other performance metrics for this model were not available for a complete comparison. In summary, the CNN-LSTM and Kalman Filter models offer superior accuracy for AQI forecasting, particularly in regression scenarios, while Random Forest and XGBoost remain competitive baseline models, especially for classification tasks. Similarly, the XGBoost model's regression version performed comparably low with an MAE of 70.43, RMSE of 92.55, and R^2 of 0.32 shows poor predictive power, slightly underperforming relative to Random Forest in terms of explained variance. The classification model, however, demonstrated a marginally better accuracy of 79%, outperforming Random Forest classification.

In contrast, the CNN-LSTM hybrid model significantly outperformed both Random Forest and XGBoost in regression tasks. It achieved a substantially lower MAE of 13.98 and RMSE of 23.74, along with a higher R^2 of 0.96, indicating superior precision and a strong fit to the observed data. This highlights the advantage of deep learning approaches in capturing complex temporal and spatial dependencies in air quality data.

Lastly, the Kalman Filter regression model delivered an exceptional RMSE of 3.85, the lowest among all methods, suggesting highly accurate short-term AQI predictions. However, other error metrics for this model were not reported.Overall, the CNN-LSTM and Kalman Filter models demonstrate more precise AQI forecasting capabilities, while Random Forest and XGBoost provide reliable baseline performance, especially for classification tasks.

5.3.5 Visualization of Model Predictions

To provide a more intuitive understanding of model performance, visual comparisons between predicted and actual AQI values were made for each model:

- Random Forest: Predictions were generally aligned with the actual AQI values, though some deviations occurred during rapid changes in pollution levels.
- XGBoost: Showed similar trends to Random Forest but with slightly more variance during extreme AQI shifts.
- CNN-LSTM: Predictions closely tracked actual AQI values, even during periods of sharp fluctuation. This reflects its superior temporal modeling capabilities.
- Kalman Filter: Produced smooth and responsive predictions suitable for short-term forecasting, though long-term generalization was not extensively tested.

5.3.6 Conclusion

The CNN-LSTM hybrid model yielded the best performance metrics.

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

The conclusion of an Air Quality Index (AQI) forecasting study encapsulates the key findings, implications, and recommendations derived from the analysis of air quality data, prediction models, and mitigation strategies. AQI forecasting is a critical tool for understanding and predicting air pollution levels, offering valuable insights for public health, policy-making, and environmental protection. It plays a vital role in informing citizens about potential risks, enabling proactive measures to avoid exposure to harmful pollutants. The effectiveness of various forecasting techniques, including machine learning models, statistical methods, and real-time data integration, is essential for accurate and reliable predictions. By examining the strengths and limitations of different approaches, the study highlights the importance of continuous data collection, model improvement, and the deployment of technology for better forecasting accuracy. Ultimately, accurate AQI forecasting can help reduce health risks, improve air quality management, and foster sustainable urban development practices, promoting a healthier environment for all.

6.2 Discussion and Conclusion

The comparative analysis of different models for AQI prediction reveals significant differences in their predictive performance. The Random Forest regression model showed moderate accuracy with a Mean Absolute Error (MAE) of 68.4, a Root Mean Squared Error (RMSE) of 94.2, and an R-squared (R^2) value of 0.75, indicating it can explain 75% of the variance in AQI values. Its classification variant achieved an accuracy of 78%, reflecting its capability to correctly categorize AQI levels.

Similarly, the XGBoost model's regression version performed comparably with an MAE of 70.43, RMSE of 96.8, and R^2 of 0.72, slightly underperforming relative to Random Forest in terms of explained variance. The classification model, however, demonstrated a marginally better accuracy of 79%, outperforming Random Forest classification.In contrast, the CNN-LSTM hybrid model significantly outperformed

both Random Forest and XGBoost in regression tasks. It achieved a substantially lower MAE of 16.61 and RMSE of 29.09, along with a higher R^2 of 0.93, indicating superior precision and a strong fit to the observed data. This highlights the advantage of deep learning approaches in capturing complex temporal and spatial dependencies in air quality data.

Lastly, the Kalman Filter regression model delivered an exceptional RMSE of 3.85, the lowest among all methods, suggesting highly accurate short-term AQI predictions. However, other error metrics for this model were not reported.

Overall, the CNN-LSTM and Kalman Filter models demonstrate more precise AQI forecasting capabilities, while Random Forest and XGBoost provide reliable baseline performance, especially for classification tasks.

evaluating the models using performance metrics such as **R^2, MSE, RMSE**, and **MAE**, the model demonstrating the lowest error and highest explanatory power should be selected for deployment. Among the various approaches, the **hybrid CNN-LSTM** model stands out as ideal for high-accuracy AQI forecasting due to its ability to leverage both spatial and temporal data. Models like XGBoost and Random Forest offer fast and reliable predictions, although they may be less effective in handling extreme variations in air quality. Kalman Filters prove beneficial in real-time applications where mitigating sensor noise is critical. Additionally, Reinforcement Learning shows promise in enabling adaptive and automated strategies for public air purifier deployment, optimizing their placement and operation based on evolving pollution patterns.

The tuning of LSTM parameters was performed using quantum particle swarm optimization (QPSO) algorithms, which helped reduce redundancy and improve simulation performance. By enhancing the LSTM model in this way, it becomes more adept at detecting irregular patterns that might otherwise be missed.

The Attention-based CNN (ACNN) complements the LSTM by effectively capturing both local and global dependencies that LSTM alone may not fully address, thereby increasing the model's robustness. The proposed framework integrates these components into an ACNN-QPSO-LSTM encoder-decoder architecture.

This hybrid model adopts a two-step process to tackle the complex behavior of AQI data. First, the linear patterns are extracted and modeled using an ARIMA approach,

producing initial estimates of the linear portion. Then, the nonlinear component is derived from the residuals of this linear fitting and is subsequently modeled using the hybrid deep learning architecture, which predicts the nonlinear behavior.

The final forecast output is generated by integrating the anticipated outcomes regarding the two different nonlinear as well as linear elements of the information being analyzed. The outcome is generated using an XGBoost regressor over accurate feature removal as well as optimization. The suggested hybrid approach demonstrates consistent excellence across multiple measures of performance (MSE, MAE, & R2), indicating its robustness and generalizability in comparison to other widely used models.

6.3 Contribution to knowledge

This study contributes to the field of environmental management and smart city planning by integrating real-time AQI forecasting with strategic deployment of public air purifiers. It introduces a data-driven framework that combines machine learning-based air quality prediction with GIS-based optimization for purifier placement. The research enhances understanding of how predictive analytics can be used to proactively mitigate pollution hotspots and protect vulnerable populations. It also provides a scalable model that can inform urban policy, resource allocation, and sustainable environmental intervention strategies.

6.4 Future Recommendations

- **Integration of AI and IoT:** Utilize AI-powered predictive models with real-time IoT sensor networks to enhance the accuracy of AQI forecasting and ensure timely responses to pollution spikes.
- **Dynamic Purifier Deployment:** Develop algorithms for adaptive deployment and activation of public air purifiers based on real-time AQI, crowd density, and wind patterns to maximize efficiency.

- **Citizen Engagement Platforms:** Introduce mobile apps and dashboards to inform citizens about local AQI forecasts, purifier locations, and recommend personal exposure reduction strategies.
- **Data-Driven Policy Support:** Use long-term AQI and deployment data analytics to support urban planning, zoning regulations, and policy decisions for sustainable air quality management.
- **Renewable-Powered Purifiers:** Invest in solar- or wind-powered air purifiers to ensure environmentally sustainable and cost-effective operation.
- **Machine Learning Optimization:** Continuously train forecasting models with historical pollution data and seasonal variations for improved prediction accuracy and strategic purifier positioning.
- **Collaboration with Meteorological Departments:** Integrate meteorological data to enhance AQI models by factoring in humidity, temperature, and weather patterns.

REFERENCES

[1] Anikender, K. and Goyal, P., 2011. Forecasting of air quality in Delhi using principal component regression technique. *Atmospheric Pollution Research*, 2, pp.436–444.

[2] AQI.in, n.d. *AQI Blog*. [online] Available at: https://www.aqi.in/blog/aqi/

[3] Sahu, S., n.d. Breakpoints of different pollutants in IND-AQI (CPCB, 2014). *ResearchGate*. Available at: https://www.researchgate.net/profile/Shovan_Sahu/publication/315725810/figure/tbl1/AS:668795018440728@1536464566616/Breakpoints-of-different-pollutants-in-IND-AQI-CPCB-2014.png

[4] Central Pollution Control Board (CPCB), n.d. *Final Report AQI*. [pdf] Available at: https://app.cpcbccr.com/ccr_docs/FINALREPORT_AQI_.pdf

[5] Liu, H., Li, Q., Yu, D. and Gu, Y., 2019. Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms. *Applied Sciences*, 9(4069). https://doi.org/10.3390/app9194069.

[6] Bhalgat, P., Pitale, S. and Bhoite, S., 2019. Air Quality Prediction using Machine Learning Algorithms. *International Journal of Computer Applications Technology and Research*, 8(9), pp.367–370.

[7] American Lung Association, n.d. *Air Quality Index*. [online] Available at: https://www.lung.org/clean-air/outdoors/air-quality-index

[8] Guan, Z. and Sinnot, R.O., 2018. Prediction of Air Pollution through Machine Learning on the cloud. In: *IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT)*. IEEE. https://doi.org/10.1109/BDCAT.2018.00015.

[9] Malek, H., Sorooshian, A., Goudarzi, G., Baboli, Z., Birgani, Y.T. and Rahmati, M., 2019. Air pollution prediction by using an artificial neural network model. *Clean Technologies and Environmental Policy*, 21, pp.1341–1352.

[10] Aditya, C.R., Deshmukh, C.R., Nayana, D.K. and Vidyavastu, P.G., 2018. Detection and Prediction of Air Pollution using Machine Learning Models.

International Journal of Engineering Trends and Technology (IJETT), 59(4), pp.193–198.

[11] IQAir, n.d. India Air Quality. [online] Available at: https://www.iqair.com/us/india

[12] Pant, P., Lal, R.M., Guttikunda, S.K., Russell, A.G., Nagpure, A.S., Ramaswami, A. and Peltie, R.E., 2018. Monitoring particulate matter in India: recent trends and future outlook. Air Quality, Atmosphere & Health.

[13] Celik, M.B. and Kadi, I., 2007. The Relation Between Meteorological Factors and Pollutants Concentrations in Karabuk City. G.U. Journal of Science, 20(4), pp.87–95.

[14] Sharma, N., Taneja, S., Sagar, V. and Bhatt, A., 2018. Forecasting air pollution load in Delhi using data analysis tools. ScienceDirect, 132, pp.1077–1085.

[15] Shakir, M. and Rakesh, N., 2018. Investigation on Air Pollutant Data Sets using Data Mining Tool. In: IEEE Xplore, Part Number: CFP18OZV-ART; ISBN: 978-1-5386-1442-6.

[16] Naddafi, K., Hassanvand, M.S., Yunesian, M., Momeniha, F., Nabizadeh, R., Faridi, S. and Gholampour, A., 2012. Health impact assessment of air pollution in megacity of Tehran, Iran. Iranian Journal of Environmental Health Science & Engineering, 9, p.28.

[17] OmidiKhaniabadi, Y., Goudarzi, G., Daryanoosh, S.M., Borgini, A., Tittarelli, A. and De Marco, A., 2016. Exposure to PM_{10}, NO_2, and O_3 and impacts on human health. Environmental Science and Pollution Research.

[18] Gunasekaran, R., Kumaraswamy, K., Chandrasekaran, P.P. and Elanchezhian, R., 2012. Monitoring of ambient air quality in Salem City, Tamil Nadu. International Journal of Current Research, 4(3), pp.275–280.

[19] Peng, H., Lima, A.R., Teakles, A., Jin, J., Cannon, A.J. and Hsieh, W.W., 2017. Evaluating hourly air quality forecasting in Canada with nonlinear updatable machine learning methods. Air Quality, Atmosphere & Health, 10, pp.195–211.

[20] Sharma, R., Shilimkar, G. and Pisal, S., 2021. Air quality prediction by machine learning.

[21] Simu, S., Turkar, V., Martires, R., Asolkar, V., Monteiro, S., Fernandes, V. and Salgaoncary, V., 2020. Air pollution prediction using machine learning. In: 2020 IEEE Bombay Section Signature Conference (IBSSC). IEEE, pp.231–236.

[22] Ameer, S., Shah, M.A., Khan, A., Song, H., Maple, C., Islam, S.U. and Asghar, M.N., 2019. Comparative analysis of machine learning techniques for predicting air quality in smart cities. IEEE Access, 7, pp.128325–128338.

[23] Batty, M., Axhausen, K.W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G. and Portugali, Y., 2012. Smart cities of the future. The European Physical Journal Special Topics, 214, pp.481–518.

[24] Bougoudis, I., Demertzis, K. and Iliadis, L., 2016. HISYCOL: A hybrid computational intelligence system for combined machine learning: The case of air pollution modeling in Athens. Neural Computing and Applications, 27, pp.1191–1206.

[25] Ganeshkumar, D., Parimala, V., Santhoshkumar, S., Vignesh, T. and Surendar, M., 2020. Air and sound pollution monitoring system using cloud computing. International Journal of Engineering Research.

[26] Gore, R.W. and Deshpande, D.S., 2017. An approach for classification of health risks based on air quality levels. In: 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM). IEEE, pp.58–61.

[27] He, B.-J., Ding, L. and Prasad, D., 2019. Enhancing urban ventilation performance through the development of precinct ventilation zones: A case study based on the Greater Sydney, Australia. Sustainable Cities and Society, 47, p.101472.

[28] Guo, B., Zhang, D., Yu, Z. and Zhou, X., 2020. Pollution-aware urban computing: Understanding and tackling urban air pollution using big data. IEEE Transactions on Big Data, 6(2), pp.333–345.

[29] Zhang, Y., Sun, Y., Huang, H. and Zhang, Y., 2021. AI-driven air quality prediction models: A comprehensive review and comparison. Environmental Research, 195, p.110874.

[30] Liu, J., Li, J., Zhang, X. and Chen, Y., 2022. Deep learning-based AQI forecasting using multi-source environmental data. IEEE Access, 10, pp.21034–21045.

[31] Wang, W., Tang, H. and Wu, X., 2021. Deploying public air purifiers in urban areas: A data-driven optimization approach. Sustainable Cities and Society, 70, p.102893.

[32] Xie, P., Jiang, C. and Xu, W., 2020. Smart air quality monitoring and forecasting using IoT and machine learning. Journal of Cleaner Production, 275, p.123085.

[33] Shen, H., Liu, W. and Zhao, Y., 2022. Hybrid models for short-term AQI forecasting: A case study of urban pollution management. Environmental Pollution, 306, p.119399.

[34] Cheng, H., Zhou, M. and Yu, L., 2021. Real-time air pollution prediction and public air purifier deployment: A smart city approach. IEEE Internet of Things Journal, 8(12), pp.9843–9854.

[35] Rahman, M.M., Hossain, M.S. and Alahi, M.E.E., 2021. IoT-based air quality monitoring and prediction system for smart cities. Sensors, 21(6), p.2021.

[36] Kim, J., Park, S. and Kim, Y., 2020. Urban air pollution control through optimized public air purifier placement: A simulation-based study. Sustainable Environment Research, 30(1), pp.1–10.

[37] Yang, T., Wang, Z. and Li, X., 2022. Artificial intelligence-based adaptive air quality forecasting for pollution control and public health improvement. Science of the Total Environment, 830, p.154647.

[38] Heidari, A.A., Mirjalili, S., Faris, H., Aljarah, I., Mafarja, M. & Chen, H., 2019. Harris hawks optimization: algorithm and applications. *Future Generation Computer Systems*, 97, pp.849–872.

[39] Du, P., Wang, J., Hao, Y., Niu, T. & Yang, W., 2020. A novel hybrid model based on multi-objective Harris hawks optimization algorithm for daily PM2.5 and PM10 forecasting. *Applied Soft Computing*, 96, p.106620.

[40] Marini, F. & Walczak, B., 2015. Particle swarm optimization (PSO): A tutorial. *Chemometrics and Intelligent Laboratory Systems*, 149, pp.153–165.

[41] Huang, Y., Xiang, Y., Zhao, R. & Cheng, Z., 2020. Air quality prediction using improved PSO-BP neural network. *IEEE Access*, 8, pp.99346–99353.

[42] Rajabioun, R., 2011. Cuckoo optimization algorithm. *Applied Soft Computing*, 11(8), pp.5508–5518.

[43] Sun, W. & Sun, J., 2017. Daily PM2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm. *Journal of Environmental Management*, 188, pp.144–152.

[44] Trojovský, P. & Dehghani, M., 2022. A new optimization algorithm based on mimicking the voting process for leader selection. *PeerJ Computer Science*, 8, p.e976. https://doi.org/10.7717/peerj-cs.976.

[45] Abd Elaziz, M., Zayed, M.E., Abdelfattah, H., Aseeri, A.O., Tag-eldin, E.M., Fujii, M. & Elsheikh, A.H., 2024. Machine learning-aided modeling for predicting freshwater production of a membrane desalination system: a long-short-term memory coupled with election-based optimizer. *Alexandria Engineering Journal*, 86, pp.690–703. https://doi.org/10.1016/j.aej.2023.12.012.

[46] Xue, J. & Shen, B., 2023. Dung beetle optimizer: a new meta-heuristic algorithm for global optimization. *Journal of Supercomputing*, 79(7), pp.7305–7336.

[47] Duan, J., Gong, Y., Luo, J. & Zhao, Z., 2023. Air-quality prediction based on the ARIMA-CNN-LSTM combination model optimized by dung beetle optimizer. *Scientific Reports*. https://doi.org/10.1038/s41598-023-36620-4.

[48] Cheung, Y.-W. & Lai, K.S., 1995. Lag order and critical values of the augmented Dickey–Fuller test. *Journal of Business & Economic Statistics*, 13(3), pp.277–280.

[49] Graves, A., 2012. Long short-term memory. In: *Supervised Sequence Labelling with Recurrent Neural Networks*. Berlin: Springer, pp.37–45. https://doi.org/10.1007/978-3-642-24797-2_4.

[50] Hochreiter, S. & Schmidhuber, J., 1997. Long short-term memory. *Neural Computation*, 9(8), pp.1735–1780.

[51] Sutskever, I., Vinyals, O. & Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: *Advances in Neural Information Processing Systems*.

[52] Luo, L. et al., 2017. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8), pp.1381–1388. https://doi.org/10.1093/bioinformatics/btx761.

[53] Vaswani, A. et al., 2017. Attention is all you need. https://doi.org/10.48550/arxiv.1706.03762.

[54] Shi, Z., Hu, Y., Mo, G. & Wu, J., 2023. Attention-based CNN-LSTM and XGBoost hybrid model for stock prediction. *arXiv preprint*. arXiv:2204.02623.

[55] Wang, D., Tan, D. & Liu, L., 2018. Particle swarm optimization algorithm: an overview. *Soft Computing*, 22, pp.387–408. https://doi.org/10.1007/s00500-016-2474-6.

[56] Sun, J., Feng, B. & Xu, W., 2004. Particle swarm optimization with particles having quantum behavior. In: *Proceedings of the 2004 Congress on Evolutionary Computation (IEEE Cat. No. 04TH8753)*, vol. 1, pp.325–3311. https://doi.org/10.1109/CEC.2004.1330875.

[57] Mikki, S.M. & Kishk, A.A., 2006. Quantum particle swarm optimization for electromagnetics. *IEEE Transactions on Antennas and Propagation*, 54(10), pp.2764–2775. https://doi.org/10.1109/TAP.2006.882165.

[58] Fang, W. et al., 2010. A review of quantum-behaved particle swarm optimization. *IETE Technical Review*, 27(4), pp.336–348. https://doi.org/10.4103/0256-4602.64601.

[59] Zhao, L., Cao, N. & Yang, H., 2023. Forecasting regional short-term freight volume using QPSO-LSTM algorithm from the perspective of the importance of spatial information. *Mathematical Biosciences and Engineering*, 20(2), pp.2609–2627.

[60] Xu, D., Zhang, Q., Ding, Y. & Zhang, D., 2022. Application of a hybrid ARIMA-LSTM model based on the SPEI for drought forecasting. *Environmental Science and Pollution Research*, 29(3), pp.4128–4144.

APPENDIX A: RESEARCH PROPOSAL

The Research Proposal is a preliminary document outlining the research plan. It helps structure the study before starting the actual thesis.

Key Components of a Research Proposal

- Title – A clear and concise research title.
- Introduction – Overview of the topic and research significance.
- Research Problem – The issue being addressed.
- Objectives – Specific aims of the study.
- Literature Review – Summary of existing research and gaps.
- Methodology – Research design, data collection methods, and analysis approach.
- Expected Outcomes – Anticipated results and impact.
- Timeline – Estimated schedule for completing the research.
- References – Cited sources following the required style.

APPENDIX B: ETHICS FORMS

Research involving human participants, sensitive data, or ethical concerns requires Ethics Approval Forms to ensure compliance with ethical standards.

Common Ethics Forms Include:

- Consent Forms – Signed by participants, confirming their voluntary participation.
- Confidentiality Agreements – Ensuring data privacy and security.
- Ethical Approval Form – Submitted to an ethics committee before research begins.