**A Project Report on**

**AIR QUALITY PREDICTION BASED ON MACHINE LEARNING**

submitted in partial fulfillment for the award of

**Bachelor of Technology**

in

**Computer Science & Engineering**

by

**CH.Naga Sai (Y20ACS422) D. Bindu Madhavi(Y20ACS411)**

**A.Meenakshi (Y20ACS407) G.V Yaswanth(Y20ACS448)**



Under the guidance of

**Prof. K. Ashok Babu.**

Department of Computer Science and Engineering

**Bapatla Engineering College**

(Autonomous)

(Affiliated to Acharya Nagarjuna University)

**BAPATLA – 522 102, Andhra Pradesh, INDIA**

**2023-2024**

**Department of**

**Computer Science & Engineering**



**CERTIFICATE**

This is to certify that the project report entitled **Air quality prediction based on machine learning** is being submitted by CH. Naga sai (Y20ACS422), D. Bindu madhavi (Y20ACS437),A. Meenakshi (Y20ACS407) and G. Venkata Yaswanth (Y20ACS448) in partial fulfillment for the award of the Degree of Bachelor of Technology in Computer Science & Engineering to the Acharya Nagarjuna University is a record of bonafide work carried out by them under our guidance and supervision.

Date:

**Signature of the Guide Signature of the HOD**

**Mr. K. Ashok Babu Mr. M. Rajesh Babu**

**Designation Prof. & Head**

**DECLARATION**

We declare that this project work is composed by ourselves, that the work contained herein is our own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

**CH. Naga Sai (Y20ACS422)**

**D. Bindu Madhavi (Y20ACS437)**

**A. Meenakshi (Y20ACS407)**

**G. V Yaswanth (Y20ACS448)**

**Acknowledgement**

We sincerely thank the following distinguished personalities who have given their advice and support for successful completion of the work.

We are deeply indebted to our most respected guide **[Mr. K. Ashok Babu, Assistant Professor],** Department of CSE, for his/her valuable and inspiring guidance, comments, suggestions and encouragement.

We extend our sincere thanks to **[Mr. M. Rajesh Babu** **]**, Prof. & Head of the Dept. for extending his cooperation and providing the required resources.

We would like to thank our beloved Principal [**Mr. Shaik Nazeer]** for providing the online resources and other facilities to carry out this work.

We would like to express our sincere thanks to our project coordinator **Dr. N. Sudhakar,** Prof. Dept. of CSE for his helpful suggestions in presenting this document.

We extend our sincere thanks to all other teaching faculty and non-teaching staff of the department, who helped directly or indirectly for their cooperation and encouragement.

**CH. Naga Sai (Y20ACS422)**

**D. Bindu Madhavi (Y20ACS437)**

**A. Meenakshi (Y20ACS407)**

**G. V Yaswanth (Y20ACS448)**

**Abstract**

Predicting air quality is a necessary step taken by governments due to its increasing impact on human health. The Air Quality Index (AQI) measures air quality, assessing pollutants like carbon dioxide, nitrogen dioxide, and carbon monoxide emitted from sources such as natural gas, coal, wood burning, industries, and vehicles. Air pollution can lead to severe diseases including lung cancer, brain diseases, and mortality. Machine learning algorithms play a crucial role in determining AQI; however, despite ongoing research, results remain somewhat inaccurate. Datasets sourced from platforms like Kaggle and air quality monitoring sites are typically divided into training and testing sets. Commonly employed machine learning algorithms for this task include Linear Regression, Decision Trees, Random Forest, Artificial Neural Networks, Support Vector Machines and XG Boost.

Efforts to enhance air quality prediction through machine learning techniques are ongoing, driven by collaboration among researchers, environmental agencies, and data scientists. The complexity of air pollution sources and pollutants poses challenges, but refining these models remains a priority for effective public health interventions. Leveraging new data sources and innovative techniques is essential for improving the accuracy of air quality forecasting. Enhanced predictions are crucial for guiding proactive policy decisions aimed at safeguarding human health.

Keywords: Air quality Index, Decision tree, Support Vector Machine, Random Forest, Linear Regression , XG Boost.

Table of Contents

[List of Figures viii](#_Toc164291514)

[List of Tables ix](#_Toc164291515)

[1 Introduction 1](#_Toc164291516)

[1.1 Introduction 1](#_Toc164291517)

[1.2 Aim of the project 3](#_Toc164291518)

[1.3 Project domain 3](#_Toc164291519)

[1.4 Scope of the project 4](#_Toc164291520)

[2 LITERATURE REVIEW 5](#_Toc164291521)

[3 PROJECT DESCRIPTION 10](#_Toc164291522)

[3.1 Introduction 10](#_Toc164291523)

[3.2 Existing System 10](#_Toc164291524)

[3.3 Proposed System 11](#_Toc164291525)

[3.4 Feasibility Study 12](#_Toc164291526)

[3.4.1 Economic Feasibility 13](#_Toc164291527)

[3.4.2 Technical Feasibility 13](#_Toc164291528)

[3.4.3 Social Feasibility 14](#_Toc164291529)

[3.5 System Specification 14](#_Toc164291530)

[3.5.1 Hardware Specification 14](#_Toc164291531)

[3.5.2 Software Specification 14](#_Toc164291532)

[4 METHODOLOGY 15](#_Toc164291533)

[4.1 Introduction 15](#_Toc164291534)

[4.2 General Architecture 16](#_Toc164291535)

[4.3 Design Phase 17](#_Toc164291536)

[4.3.1 Data Flow Diagram 17](#_Toc164291537)

[4.3.2 Use Case Diagram 18](#_Toc164291538)

[4.3.3 Class Diagram 20](#_Toc164291539)

[4.3.4 Sequence Diagram 21](#_Toc164291541)

[4.3.5 Collaboration Diagram 22](#_Toc164291542)

[4.3.6 Activity Diagram 23](#_Toc164291543)

[4.4 Module Description 24](#_Toc164291544)

[4.4.1 Data Preprocessing 24](#_Toc164291545)

[4.4.2 Feature Selection 25](#_Toc164291546)

[4.4.3 Modeling 25](#_Toc164291547)

[4.5 Steps to implement the project 26](#_Toc164291548)

[4.5.1 Installation of anaconda software 26](#_Toc164291549)

[4.5.2 Installation of packages 26](#_Toc164291550)

[4.5.3 Activate new environment 27](#_Toc164291551)

[5 IMPLEMENTATION 28](#_Toc164291552)

[5.1 Input and Output 28](#_Toc164291553)

[5.1.1 Input Design 28](#_Toc164291554)

[5.1.2 Output Design 29](#_Toc164291555)

[6 RESULTS AND DISCUSSION 32](#_Toc164291556)

[6.1 Efficiency of Proposed System 32](#_Toc164291557)

[6.2 Comparison of Existing and Proposed system 33](#_Toc164291558)

[7 CONCLUSION AND FUTURE ENHANCEMENTS 34](#_Toc164291559)

[7.1 Conclusion 34](#_Toc164291560)

[7.2 Future Enhancements 35](#_Toc164291561)

[8 References 37](#_Toc164291562)

List of Figures

[**Figure 4.1 General Architecture** 16](#_Toc164147257)

[**Figure 4.2 Data Flow Diagram Level 0** 16](#_Toc164147258)

[**Figure 4.3 Data Flow Diagram Level 1** 17](#_Toc164147259)

[**Figure 4.4 Use Case Diagram** 18](#_Toc164147260)

[**Figure 4.5 Class Diagram** 19](#_Toc164147261)

[**Figure 4.6 Sequence Diagram** 20](#_Toc164147262)

[**Figure 4.7 Collaboration Diagram** 21](#_Toc164147263)

[**Figure 4.8 Activity Diagram** 22](#_Toc164147264)

List of Tables

[Table 3.1 12](#_Toc164147313)

# Introduction

## Introduction

The impact of machine learning technologies on nearly every industry is becoming increasingly evident, and air pollution prediction is no exception. In the context of smart cities, Air pollution is a growing concern in megacities across the world, with adverse effects on human health and the environment. Historical data can be used to train machine learning models, allowing them to learn from their mistakes and ad vance over time. The models may grow even more precise and dependable as more data is gathered. Machine learning algorithms for predicting real-time air quality have a wide range of uses, such as early alerts for excessive pollution levels, guidance for traffic and urban planning decisions, and informing public health policy. Machine learning can assist us in making well-informed decisions to safeguard our health and the environment by precisely forecasting air quality levels.

Real-time air quality prediction using advanced machine learning (ML) techniques can help mitigate the harmful effects of air pollution. This approach involves analyzing large volumes of real-time data from various sources, such as air quality sensors, weather stations, and satellite imagery, to accurately predict air quality levels in a particular area. The use of machine learning algorithms to handle massive amounts of data from numerous sources, such as air quality sensors, weather information, and satellite images, is necessary for real-time air quality prediction. In order to anticipate the levels of air quality, the computers analyze the data to find patterns and correlations. Advanced ML techniques such as deep learning, ensemble models, and hybrid algorithms are used to develop highly accurate predictive models that can forecast air quality levels hours, days, or even weeks in advance. The most relevant papers were chosen using a combination of searching the most prominent databases and applying the appropriate filters.

These models can be trained on historical data to learn the patterns and correlations between different environmental factors and air pollution levels, and then used to forecast future pollution levels in real-time. Instead of using basic machine learning techniques, authors now employ more complex ones. The output from these models can be used to inform decision-making processes and help policymakers take targeted actions to improve air quality. These models can also identify the main contributors to air pollution, such as industrial emissions, traffic congestion, and natural events like wildfires. Real-time air quality prediction can assist policymakers in making informed decisions about environmental policies, such as implementing emission control measures, improving urban planning, and monitoring the impact of policy interventions. The use of advanced ML techniques in air quality prediction represents a significant step towards improving the health and wellbeing of citizens in megacities worldwide.

## Aim of the project

* The objective of the system is to construct a system that can capture the fluctuation in air quality that is induced due to changes in the meteorological conditions in order to offer short-term, real-time predictions of the air quality index cities in India.
* The purpose is to make air quality predictions in real-time using a model trained with pollutant concentration data.

## Project domain

The project domain for real-time air quality prediction using machine learning falls under the domain of environmental technology. The project aims to provide real-time air quality predictions using machine learning algorithms to analyse environmental data. The system will use various sensors to collect air quality data, preprocess the data, and use machine learning algorithms to predict future air quality levels. The project will be beneficial to various industries, including transportation, manufacturing, and agriculture, by providing more accurate air quality information to facilitate better decision-making. The project could also help to reduce the risks associated with poor air quality and improve public health by enabling individuals to take appropriate precautions.

## Scope of the project

This model can handle nonlinear emission levels, atmospheric science, and meteorological relationships. This model also helps us identify air-quality drivers and evaluate future traffic-emission controls. We prove that fleet electrification cannot fully benefit from local vehicle emissions mitigation. Off-road and volatile chemical emissions must be strictly regulated to improve Indian air quality.

The project would involve applying advanced ML techniques such as deep learning, neural networks, and regression analysis to analyse the data and develop a predictive model. The model would be trained on historical data and then used to forecast air quality levels for future time periods. The predictions could be visualized using graphs or maps to help decision-makers and the public understand the air quality situation in megacities.

The project would also require considering the impact of various factors on air quality, such as traffic patterns, industrial activity, and weather conditions, to ensure the model provides accurate predictions. The ultimate goal would be to create a tool that can be used to inform policymakers and the public about air quality levels, allowing for more informed decision-making and targeted interventions to improve air quality in megacities.

# LITERATURE REVIEW

[1] Temesegan Walelign Ayele, et al, “Air pollution monitoring and prediction using IoT”, Second International Conference on Inventive Communication and Computa- tional Technologies (ICICCT), 2018. This paper proposed an IoT-based air pollution monitoring and prediction system.The model is improved in two ways based on an existing method: the feature attribute value and the weighting of the information gain. Accuracy and computational complexity are both enhanced. This system can be used to monitor air pollutants in a specific area, perform air quality analysis, and forecast air quality. The proposed system will monitor air pollutants by combining IoT with a machine learning algorithm known as Recurrent Neural Network, more specifically Long Short Term Memory (LSTM).

[2] Saba Ameer, et al, “Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities”, IEEE Access ( Volume: 7), 2019. This pa- per proposed four advanced regression techniques to predict pollution and present a comparative study to determine the best model for accurately predicting air quality in terms of data size and processing time. The researchers conducted experiments with Apache Spark and estimated pollution using multiple datasets. For the comparison of these regression models, the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were used as evaluation criteria.

[3] Yi-Ting Tsai, et al, “Air Pollution Forecasting Using RNN with LSTM”, IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, 2018. This paper proposed a method for forecasting PM2.5 concentrations that combine RNN (Recurrent Neu- ral Network) and LSTM (Long Short-Term Memory). The researchers use Keras, a Python-based high-level neural networks API that can run on top of Tensorflow, to build a neural network and run RNN with LSTM through Tensorflow.

[4] Venkat Rao Pasupuleti, et al, “Air Quality Prediction Of Data Log By Machine Learning”, 6th International Conference on Advanced Computing and Communica- tion Systems (ICACCS), 2020. With advances in machine learning technology, it is now possible to predict pollutants based on historical data. This paper introduced a device that can take current pollutants and, with the help of past pollutants, run an algorithm based on machine learning to predict future pollutant data. The sensed data is saved in an Excel sheet for later analysis. These sensors are used to collect pollutant data on the Arduino Uno platform.

[5] Shengdong Du, et al, “Deep Air Quality Forecasting Using Hybrid Deep Learn- ing Framework”, Transactions on Knowledge and Data Engineering (Volume: 33, Issue: 6), 2021. This paper proposed a deep learning model for air quality (pri- marily PM2.5) forecasting that uses a hybrid deep learning architecture to learn the spatial-temporal correlation features and interdependence of multivariate air quality- related time series data. Because multivariate air quality time series data is nonlinear and dynamic, the base modules of the model include one-dimensional Convolutional Neural Networks (1D-CNNs) and bi-directional Long Short-term Memory networks (Bi-LSTM). The former extracts local trend and spatial correlation feature, while the latter learns spatial-temporal dependencies. Then, for shared representation features learning of multivariate air quality-related time series data, the researchers design a jointly hybrid deep learning framework based on one-dimensional CNNs and Bi- LSTM.

[6] Ke Gu, et al, “Recurrent Air Quality Predictor Based on Meteorology - and Pollution-Related Factors”, IEEE Transactions on Industrial Informatics ( Volume: 14, Issue: 9), 2018. This paper proposed a heuristic recurrent air quality predic- tor (RAQP) to infer air quality in this paper. The RAQP uses key meteorological and pollution-related variables to calculate air pollutant concentrations (APCs), such as fine particulate matter (PM2.5). The RAQP method repeatedly employs the 1-h prediction model, which learns current records of meteorology and pollution-related factors in order to predict air quality 1 hour later, and then estimates air quality after several hours. Extensive experiments show that the RAQP predictor outperforms rel- evant state-of-the-art techniques and nonrecurrent methods for air quality prediction.

[7] Bo Liu, et al, “A Sequence-to-Sequence Air Quality Predictor Based on the n-Step Recurrent Prediction”, IEEE Access (Volume: 7), 2019. Using Beijing as an example, this study proposed an attention-based air quality predictor (AAQP) to better protect people from air pollution. The AAQP is a seq2seq model that uses his- torical air quality data as well as weather data to forecast future air quality indexes. The experimental results confirmed that the AAQP with n-step recurrent prediction outperformed the related arts because the error accumulation was reduced and the training time was significantly reduced when compared to the original seq2seq at- tention model.

[8] Baowei Wang, et al, “Air Quality Forecasting Based on Gated Recurrent Long Short Term Memory Model in Internet of Things”, IEEE Access (Volume: 7), 2019. This model is an improvement and enhancement of the existing Long Short Term Memory prediction method (LSTM). The experiment combines data from the IoT node and information from the national environmental protection department. First, 96 consecutive hours of data from four cities were chosen as experimental samples. The experimental results are nearly identical to the true value. Then, as a train and test dataset, the researchers chose daily smog data from 2014/1/1 to 2018/1/1. It in- cludes smog data for 74 cities. The first 70percent of the data was used for training, while the remainder was used for testing. The results of this experiment show that this model can predict better.

[9] Yuanni Wang, et al, “Air Quality Predictive Modeling Based on an Improved De- cision Tree in a Weather-Smart Grid”, IEEE Access (Volume: 7), 2019. This paper proposed an improved decision tree method to improve the time performance and accuracy of prediction with a large amount of data. The model is improved in two ways based on an existing method: the feature attribute value and the weighting of the information gain. Accuracy and computational complexity are both enhanced. The experimental results show that the improved model outperforms the traditional methods in terms of accuracy and computational complexity. Furthermore, it is more efficient in dealing with classification and prediction with large amounts of air qual- ity data. Furthermore, it is capable of making accurate predictions for future data.

[10] Van-Duc Le, et al, “Spatiotemporal Deep Learning Model for Citywide Air Pollution Interpolation and Prediction”, IEEE International Conference on Big Data and Smart Computing (BigComp), 2020. This paper proposed the use of the Convo- lutional Long Short-Term Memory (ConvLSTM) model, a hybrid of Convolutional Neural Networks and Long Short-Term Memory that automatically manipulates both spatial and temporal data features. In particular, the researchers show how to convert air pollution data into image sequences that use the ConvLSTM model to interpolate and predict air quality for the entire city at the same time. Also, they demonstrate that their approach is applicable to spatiotemporal air pollution problems and outper- forms previous research in this area .

# PROJECT DESCRIPTION

## Introduction

The project description section in a project report succinctly outlines the objectives, scope of work, methodology, key deliverables, and constraints of the initiative. It provides a clear roadmap for understanding the project's purpose, goals, and intended outcomes. This section is crucial for stakeholders to grasp the project's focus and approach quickly. By concisely summarizing these key aspects, the project description sets the stage for the detailed analysis and findings presented in the report.

## Existing System

One existing system that uses XG Boost for real-time air quality prediction in megacities is ”Air Quality Prediction using using Machine Learning Algorithems”. While XG Boost is a widely used machine learning algorithm that can perform well in many contexts, it also has some disadvantages in the context of air quality prediction. Here are a few:

**Assumption of Independence:** XG Boost assumes that the features used for classification are independent of each other. However, in the context of air quality prediction, this assumption may not hold true, as air quality is influenced by complex interactions between various environmental factors.

**Limited Expressiveness:** XG Boost has limited expressiveness, which can limit its ability to capture complex relationships and patterns in the data.

**Sensitivity to Outliers:** XG Boost is sensitive to outliers and noise in the data, which can negatively impact its performance.

**Limited Robustness** XG Boost can be highly sensitive to changes in the data distribution, which can reduce the robustness of the algorithm in real-world scenarios.

Despite these disadvantages, XG Boost can still be effective for air quality pre- diction when used appropriately and with careful consideration of their limitations.

## Proposed System

A proposed system for real-time air quality prediction in megacities using Cat Boost algorithm can provide several advantages over other methods.

Here are a few:

**Robustness:** Cat Boost algorithm are robust to changes in the data distribution, making them effective in real-world scenarios where data may be variable.

**High accuracy:** Cat Boost algorithm are known to have high accuracy in classification tasks, which is critical for accurate air qual- ity prediction.

**Feature selection** Cat Boost algorithm can be used for feature selection, which can help identify the most important features for air quality prediction and streamline the prediction process.

**Real-time prediction:** Cat Boost algorithm can make real-time predictions, which is critical for air quality prediction in megacities, where timely alerts can help prevent negative health outcomes.

Overall, the Cat Boost algorithm offer several advantages, including robustness, non-parametric flexibility, interpretability, high accuracy, feature selection, and real-time prediction, making them promising approaches for real-time air quality prediction in mega cities. It can be seen in Fig. 3.1 that the standard limits of pollutants are shown for a common residential area. In situations where there are vast differences between the ideal parameters and the actual values, hazardous air quality occurs as a result. This air pollution can cause health problems such as respiratory illnesses, eye and skin irritations, and can even lead to heart attacks and cancer.

Table 3.1

|  |  |
| --- | --- |
| **Pollutant** | **Standard concentration in ambient air** |
| PM2.5 | 60 |
| PM10 | 100 |
| SO2 | 90 |
| NO2 | 80 |
| CO | 04 |
| O3 | 180 |

## Feasibility Study

The purpose of this feasibility study is to evaluate the practicality and viability of the Real-Time Air Quality Prediction using Advanced Machine Learning (ML) Tech- niques project. This project aims to develop a real-time air quality prediction system that uses advanced ML techniques to provide accurate and reliable air quality pre- dictions to individuals and organizations.

### Economic Feasibility

The economic feasibility of the project will depend on the cost of hardware, soft- ware, and labor required to develop and implement the real-time air quality prediction system. Additionally, the project will require ongoing maintenance and updates to ensure the accuracy and reliability of the system. To ensure economic feasibility, the project team will need to carefully evaluate and select cost-effective hardware and software solutions. Additionally, the team will need to consider the ongoing maintenance costs and develop a sustainable business model that can support the ongoing costs of the system.

### Technical Feasibility

To develop a real-time air quality prediction system, the project will require access to historical air quality data and real-time air quality monitoring data. Additionally, the project will need to incorporate advanced ML techniques, such as deep learning and neural networks, to accurately predict air quality levels. The feasibility of the technical aspect of the project will depend on the availability of data sources and the availability of expertise in advanced ML techniques. There are many publicly available data sources for air quality monitoring that can be used for the project. Furthermore, there are many ML experts available who can assist in developing and implementing the ML techniques required for the project.

### Social Feasibility

The social feasibility of the Real-Time Air Quality Prediction using Advanced ML Techniques project will depend on its potential impact on society and the acceptance of the system by individuals and organizations. The project has the potential to benefit society by providing individuals and organizations with real-time air quality predictions that can help them make informed decisions about their activities and health. Additionally, the project can help raise awareness about the importance of air quality monitoring and encourage individuals and organizations to take steps to reduce air pollution.

## System Specification

### Hardware Specification

Processor : Ryzen 7

Hard disk : 512 GB

RAM : 4GB(minimum)

Keyboard : 110 keys enhanced

### Software Specification

Operating system : Windows 11

Language : Python

# METHODOLOGY

## Introduction

The methodology section of this project report provides a detailed account of how we executed the project objectives from start to finish. To begin with, we carefully outlined the research design, which included selecting appropriate methods for data collection and analysis based on the project's goals. We employed a combination of quantitative surveys and qualitative interviews to gather primary data from target participants. The survey questions were designed to capture specific information related to our research questions, while the interviews provided deeper insights into participants' perspectives and experiences. Additionally, we utilized existing literature and secondary sources to contextualize our findings and validate our analysis.

Furthermore, this methodology section elucidates the steps taken to ensure the reliability and validity of our research. By meticulously documenting our methodology, we aim to provide transparency and enable other researchers to replicate or build upon our work effectively. This section underscores our commitment to rigor and integrity in conducting the project and interpreting its outcomes.

## General Architecture

A diagram of a process flow

Description automatically generated

**Figure 4.1 General Architecture**

**Figure 4.1** shows the architectural design for advanced ML-based real-time air quality prediction in megacities would include a number of different parts. A sensor network would be used to gather data on air quality, a database would be used to store the data, a machine learning module would be used to analyze the data, and a visualization system would be used to provide the predictions.

## Design Phase

### A diagram of air quality prediction Description automatically generatedData Flow Diagram

**Figure 4.2 Data Flow Diagram Level 0**

A diagram of a model

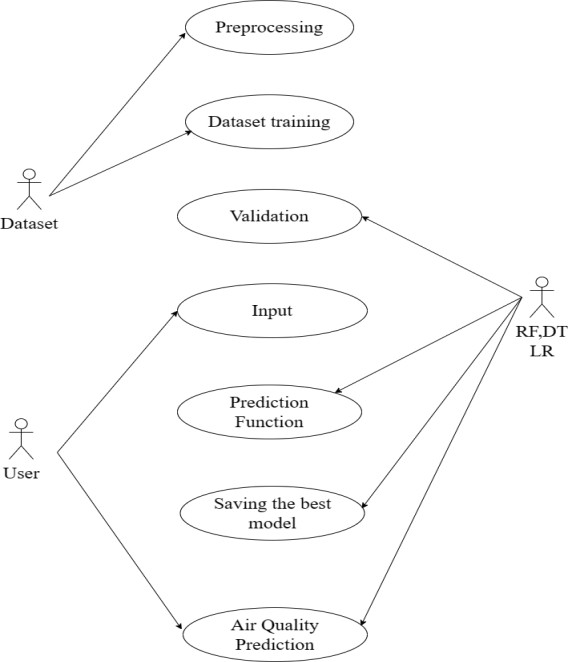
Description automatically generated

**Figure 4.3 Data Flow Diagram Level 1**

In **Figure 4.3** the data flow diagram would show how data would move between the various system components. Typically, the sensor network, database, machine learn- ing module, and visualization system are shown in the diagram. Data on air quality would be gathered by the sensor network and sent to the database.

The data would be analyzed by the machine learning module, which would then produce predictions. These predictions would then be passed to the visualization system for presentation.

### Use Case Diagram



**Figure 4.4 Use Case Diagram**

The use case diagram in **Figure 4.4**, would show the various ways in which a user and a system might interact. Typically, the diagram would show various use scenarios, such as requesting air quality data, utilizing machine learning algorithms to analyze the data, and visualizing the results.

**Figure 4.4**, is a graphical depiction of a user’s possible interaction with the system. The use case involves pre-processing which in result removes the noisy data. After pre-processing, datasets are used to train the ML models using historical data and advanced ML techniques. Making real-time predictions using the trained models based on new data collected and feature engineering.

### Class Diagram

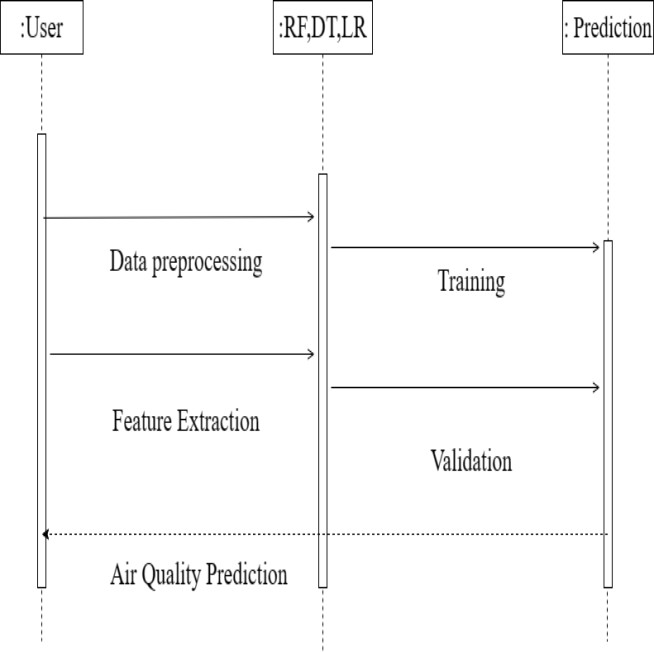
### 

**Figure 4.5 Class Diagram**

In general, class diagram in UML indicates a type of static structure diagram that de- scribes the structure of the model by describing the system’s classes, attributes,entities and their relationships among the members.

The class diagram in **Figure 4.5**, would show the many classes and how they relate to one another inside the system. Classes like the Sensor class, the Database class, the Machine Learning Module class, and the Visualization System class would normally be included in the diagram.

### Sequence Diagram



**Figure 4.6 Sequence Diagram**

In **Figure 4.6**, the process is showed in form of interactions in time sequence in field of engineering. The user-interface is the front-end of the system, it receives inputs from the user and displays the output of the prediction system. The input data is preprocessed and cleaned to remove any invalid or missing values. The model then processes the data and generates a prediction of air quality index for each location in the mega-cities.

The sequence diagram in Fig **Figure 4.6**, would show how the various parts of the sys- tem interacted over time. The sensor network, the database, the machine learning module, and the visualization system would normally be included as players in the diagram.

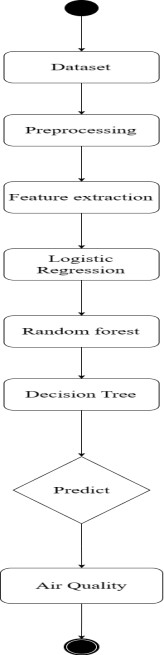
### Collaboration DiagramA diagram of a data processing process Description automatically generated

**Figure 4.7 Collaboration Diagram**

**Figure 4.7**, describes the Collaboration diagram, also known as a communication dia- gram, that illustrates the interactions and relationships among in a system. Collabo- ration diagrams visualize the flow of messages exchanged between the objects in a system and the order in which these messages are sent.

The collaboration diagram in **Figure 4.7**, would show the interactions and connections among the various system components. The sensor network, the database, the ma- chine learning module, and the visualization system would normally be included as players in the diagram.

### Activity Diagram



**Figure 4.8 Activity Diagram**

In **Figure 4.8**, the activity diagram shows the system’s action and decision making flow. The system receives the information from multiple sources, such as weather stations, traffic sensors, and air quality sensors. Th air quality index values are stored in a database or data warehouse for further analysis and visualization.

## Module Description

### Data Preprocessing

* Data extraction and preprocessing module is the initial step in the real-time air quality prediction system using Decision tree algorithm. In this module the data is extracted from the websites and converts it into an HTML file for exploratory data analysis(EDA).
* The next step is data cleaning, where noise is removed from the dataset and features are normalized to bring all the values to a single scale for better perfor- mance of the model.
* This is an essential step as dataset has varying scale values, such as single digit, double digit and triple digit values.
* After data cleaning, EDA is performed on the preprocessed data.

This module is crucial as it ensures that the dataset is ready for the subsequent modules for real-time air quality prediction.

### Feature Selection

* After Data preprocessing the focus is on feature selection to improve the performance of the model.
* Among all the available features, the ones with highest importance are being selected. In this model PM2.5,PM10,SO2,NO2,O3,CO are selected as the most important features that impact the decision.
* This avoids the overfitting problem by avoiding or reducing irrelevant or partially relevant features in the dataset.
* The availability of many features may arise the problem of curse of dimensionality, that may in turn reduce the efficiency of the model by a greater extent.
* This step is followed by splitting the dataset into training and testing set and training the model with the help of training data set.

### Modeling

* As part of this step, a model is built, which is then trained on the training data in order for it to gain experience and then be able to predict future results based on the gained experience.
* As a first step, we will be using the Decision Tree algorithm in order to calculate the air quality index with the help of the features that were selected during the feature selection process.
* As a result of this regression step, the air quality index is used to categorize the air quality into different categories based on the air quality index that was obtained.
* To calculate the accuracy score of the models, different metrics are used, and the one with the highest accuracy is used.

## Steps to implement the project

### Installation of anaconda software

Download install the Anaconda package 64-bit version and choose the Python 3.11 version.

### Installation of packages

Install the necessary packages that are mentioned in the requirements.txt file.

Open your anaconda prompt and clone the repository git clone If you have not al- ready created a new virtual environment in Step 1, then create a conda environment conda create -n your env name python=3.11.

### Activate new environment

Activate the new environment using the anaconda prompt. activate your env name python setup.py build ext –in place or try the following as an alternative pip install–e

# IMPLEMENTATION

## Input and Output

### Input Design

The input design for real-time air quality prediction using machine learning would involve the following:

**Environmental Data:** The primary input for the system would be environmental data, including data from various sensors and other sources. This data would include information about air quality, such as the concentration of pollutants, temperature, humidity, and wind speed.

**Data Preprocessing:** The system would require data preprocessing, including data cleaning, data transformation, and feature engineering, to prepare the data for analysis.

**Machine Learning Algorithms:** The system would use machine learning algo- rithms, such as regression algorithms, random forest algorithm, and convolutional neural networks, to predict future air quality levels based on the input data.

Overall, the input design would need to support environmental data input, data pre- processing, time-series data analysis, machine learning algorithms for air quality prediction, and ensemble learning techniques. The system should be able to process environmental data in real-time and predict air quality levels as quickly as possible to facilitate better decision-making.

### Output Design

The output design for real-time air quality prediction using machine learning would involve the following:

**Air Quality Predictions:** The primary output of the system would be air quality predictions. The system would use machine learning algorithms to predict future air quality levels based on the input data. The output would include information about the concentration of pollutants, temperature, humidity, and wind speed.

**User Interface:** The system would need to have a user-friendly interface for displaying the air quality predictions to the users. The interface should be designed to be easily understandable and intuitive.

**Data Visualization:** The system could provide data visualization to help users understand the data and identify patterns and trends in the data. This could help to improve the accuracy of the system and ensure that it is providing effective air qual- ity predictions.

**Historical Data:** The system could also provide historical air quality data to help users understand past air quality trends and make informed decisions about fu- ture activities.

Overall, the output design would need to support real-time air quality predictions, a user-friendly interface for displaying the predictions, an alert mechanism, data vi- sualization, and historical data. The system could help individuals and industries make informed decisions about outdoor activities and take appropriate precautions to protect their health.

A screenshot of a login form

Description automatically generated

A screen shot of a login form

Description automatically generated

A screen shot of a computer

Description automatically generated

A black square with white text and numbers on it

Description automatically generated

A black square with white text

Description automatically generated

# RESULTS AND DISCUSSION

## Efficiency of Proposed System

The proposed system for real-time air quality prediction using machine learning (ML) can be efficient in providing accurate and timely predictions of air quality levels. The system employs machine learning algorithms to analyze data from air quality sensors and predict future air quality levels, enabling users to take appropriate measures to reduce their exposure to pollutants.

The efficiency of the system can be evaluated based on its accuracy, speed, and reliability in predicting air quality levels. The system should be able to accurately predict air quality levels based on real-time data from sensors, while minimizing false positives and negatives.

The system’s accuracy can be improved through the use of advanced machine learning algorithms and data preprocessing techniques. Additionally, the use of multiple sensors and data sources can enhance the system’s accuracy and effectiveness.

The system’s speed and efficiency can be improved through the use of optimized algorithms and hardware acceleration techniques. These optimizations can enable the system to provide real-time predictions of air quality levels, allowing users to take appropriate actions to protect their health.

Overall, the proposed system for real-time air quality prediction using ML has the potential to provide an efficient and effective way to monitor and predict air quality levels, improving public health and reducing the risks associated with air pollution.

## Comparison of Existing and Proposed system

Existing systems for real-time air quality prediction typically rely on traditional statistical models and are limited in their accuracy and reliability. These systems may also require extensive manual intervention for calibration and data processing.

In contrast, the proposed system for real-time air quality prediction using ML employs advanced machine learning algorithms to improve accuracy and reliability. The system can predict air quality levels in real-time based on data from multiple sensors, allowing for more precise and accurate predictions.

The proposed system also has the potential to adapt to changes in air quality patterns and provide more accurate predictions over time. Additionally, the system can be customized to provide personalized air quality predictions based on user specific data, such as location and activity level. In terms of speed, the proposed system is designed to provide real-time predictions of air quality levels, enabling users to take immediate actions to reduce their exposure to pollutants. The system’s use of ML algorithms allows for fast and efficient processing of large amounts of data, making it more reliable and efficient than existing systems.

Overall, the proposed system for real-time air quality prediction using ML has several advantages over existing systems, including improved accuracy, reliability, and speed. These advantages can lead to more effective measures to protect public health and reduce the risks associated with air pollution.

# CONCLUSION AND FUTURE ENHANCEMENTS

## Conclusion

The purpose of this study is to survey the most up-to-date literature on the topic of air quality prediction in order to provide a broad overview of the many methodologies that have been proposed. Since air quality prediction is a broad subject, we have established some guidelines in order to zero in on a specific objective. The databases Scopus and IEEE Xplore were queried with a predefined query to find relevant documents. We chose research published after 2002, removing irrelevant papers using inclusion/exclusion criteria, for further investigation. In the end, 41 manuscripts made the cut. From our analysis of these articles, we have drawn out their most salient characteristics, and we have used these conclusions to draw connections between and compare the various works. When geography is considered, the findings reveal that 26 of the studies used China as a case study. The key discovery was that including a dataset of additional elements that affect air quality in addition to air quality data is useful for increasing the accuracy of air quality prediction.

## Future Enhancements

Future enhancements for real-time air quality prediction using ML could include :

**Integration with other data sources:** The system could be enhanced to integrate data from other sources, such as weather forecasts or traffic patterns, to provide more accurate predictions and insights.

**Personalized predictions:** The system could be enhanced to provide personalized predictions based on user-specific data, such as location, activity level, and health status, to enable individuals to take appropriate measures to protect their health.

**Advanced data preprocessing techniques:** The system could be enhanced to in- corporate advanced data preprocessing techniques, such as data cleaning and feature engineering, to improve the accuracy and reliability of the ML models.

**Multi-modal sensing:** The system could be enhanced to incorporate data from mul- tiple sensors, such as air quality sensors and wearable sensors, to provide a more comprehensive and accurate view of individual exposure to pollutants.

**Integration with smart city infrastructure:** The system could be enhanced to in- tegrate with smart city infrastructure, such as traffic management systems or urban planning data, to provide more insights and recommendations for improving air qual- ity.

**Improved explainability:** The system could be enhanced to provide more trans- parency and explainability in its predictions, allowing users to understand the factors that contribute to air quality levels and make more informed decisions

Overall, these enhancements can improve the effectiveness and functionality of the system, enabling it to provide more accurate and reliable predictions of air quality levels and more tailored recommendations for individuals and communities to im- prove air quality.

# References

[1] G. B. Fioccola, R. Sommese, I. Tufano, R. Canonico, and G. Ventre, “Polluino: An Efficient Cloud-based Management of IoT Devices for Air Quality Monitor- ing”, in Proc. IEEE Research and Technologies for Society and Industry Lever- aging a better tomorrow, pp.1–6, Sep. 2016.

[2] T. M. Chiwewe and J. Ditsela, “Machine Learning Based Estimation of Ozone Using Spatio-Temporal Data from Air Quality Monitoring Stations,” INDIN 2016 Proc. 14th IEEE Int. Conf.,pp. 58–63, 2016.

[3] C.-y. Chiu and Z. Zhang, “The air quality evaluation based on gas sensor array,” in 2017 China Semiconductor Technology International Conference (CSTIC). IEEE, 2017, pp. 1–5.

[4] S. Kumar and A. Jasuja, “Air quality monitoring system based on iot using rasp- berry pi,” in 2017 International Conference on Computing,Communication and Automation (ICCCA). IEEE, 2017, pp. 1341–1346.

[5] You-Chiun Wang, Guan-Wei Chen, “ Efficient Data Gathering and Estima- tion for Metropolitan Air Quality Monitoring by Using Vehicular Sensor Net- works”,IEEE Transactions On Vehicular Technology, 2017

[6] Temesegan Walelign Ayele, Rutvik Mehta, “Air pollution monitoring and predic- tion using IoT”, Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018

[7] Z. Hu, Z. Bai, K. Bian, T. Wang, and L. Song, “Real- time Fine-grained Air Qual- ity Sensing Networks in Smart City: Design, Implementation and Optimization”, arXiv preprint arXiv:1810.08514, Oct. 2018.

[8] Yi-Ting Tsai, Yu-Ren Zeng, Yue-Shan Chang, “Air Pollution Forecasting Using RNN with LSTM”, IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technol- ogy Congress, 2018

[9] Ke Gu, Junfei Qiao, Weisi Lin, “Recurrent Air Quality Predictor Based on Meteorology- and Pollution-Related Factors”, IEEE Transactions on Industrial Informatics ( Volume: 14, Issue: 9), 2018

[10] Saba Ameer, Munam Ali Shah, Abid Khan, Houbing Song, Carsten Maple, Saif Ul Islam, Muhammad Nabeel Asghar, “Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities”, IEEE Access (

Volume: 7), 2019

[11] Bo Liu, Shuo Yan, Jianqiang Li, Guangzhi Qu, Yong Li, Jianlei Lang, Rentao Gu, “A Sequence-to-Sequence Air Quality Predictor Based on the n-Step Recur- rent Prediction”, IEEE Access (Volume: 7), 2019

[12] Baowei Wang, Weiwen Kong, Hui Guan, Neal N. Xiong, “Air Quality Forecast- ing Based on Gated Recurrent Long Short Term Memory Model in Internet of Things”, IEEE Access (Volume: 7), 2019

[13] Yuanni Wang, Tao Kong, “Air Quality Predictive Modeling Based on an Im- proved Decision Tree in a Weather-Smart Grid”, IEEE Access (Volume: 7), 2019

[14] Venkat Rao Pasupuleti, Uhasri, Pavan Kalyan, Srikanth, Hari Kiran Reddy, “Air Quality Prediction Of Data Log By Machine Learning”, 6th International Con- ference on Advanced Computing and Communication Systems (ICACCS), 2020

[15] Van-Duc Le, Tien-Cuong Bui, Sang-Kyun Cha, “Spatiotemporal Deep Learning Model for Citywide Air Pollution Interpolation and Prediction”, IEEE Interna- tional Conference on Big Data and Smart Computing (BigComp), 2020

[16] Shengdong Du, Tianrui Li, Yan Yang, Shi-Jinn Horng, “Deep Air Quality Fore- casting Using Hybrid Deep Learning Framework”, Transactions on Knowledge and Data Engineering (Volume: 33, Issue: 6), 2021’