RAJAGIRI SCHOOL OF ENGINEERING & TECHNOLOGY (AUTONOMOUS)

*Project Phase II Report On*

# AirGuardian: Air Quality Analysis and Disease Warning System

*Submitted in partial fulfillment of the requirements for the award of the degree of*

## Bachelor of Technology

*in*

## *Computer Science and Engineering*

**By**

**Elizabath Gigi (U2003077)**
**Govind Prasad (U2003089)**
**Manu Jaison (U2003126)**
**Naina Fathima K. (U2003140)**

**Under the guidance of**
**Dr. Jisha G.**

**Department of Computer Science and Engineering**
**Rajagiri School of Engineering & Technology (Autonomous)**
**(Parent University: APJ Abdul Kalam Technological University)**
**Rajagiri Valley, Kakkanad, Kochi, 682039**
**May 2024**

# CERTIFICATE

*This is to certify that the project report entitled* ***"AirGuardian: Air Quality Analysis and Disease Warning System"*** *is a bonafide record of the work done by* ***Ms. Elizabath Gigi (U2003077), Mr. Govind Prasad (U2003089), Mr. Manu Jaison (U2003126), Ms. Naina Fathima K. (U2003140)****,submitted to the Rajagiri School of Engineering & Technology (RSET) (Autonomous) in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (B. Tech.) in Computer Science and Engineering during the academic year 2023-2024.*

**Dr. Jisha G.**
Project Guide
Associate Professor
Dept. of CSE
RSET

**Dr. Tripti C.**
Project Coordinator
Associate Professor
Dept. of CSE
RSET

**Dr. Preetha K. G.**
Head of the Department
Professor
Dept. of CSE
RSET

# ACKNOWLEDGMENT

We wish to express our sincere gratitude towards **Dr. P. S. Sreejith**, Principal of RSET, and **Dr. Preetha K. G.**, Head of the Department of Computer Science and Engineering for providing us with the opportunity to undertake our project, **AirGuardian: Air Quality Analysis for Disease Prediction**.

We are highly indebted to our project coordinators, **Dr. Tripti C.**, Associate Professor, Department of CSE, for their valuable support.

It is indeed our pleasure and a moment of satisfaction for us to express our sincere gratitude to our project guide **Dr. Jisha G.**, Associate Professor, Department of CSE, for her patience and all the priceless advice and wisdom she has shared with us.

Last but not the least, we would like to express our sincere gratitude towards all other teachers and friends for their continuous support and constructive ideas.

**Elizabath Gigi**

**Govind Prasad**

**Manu Jaison**

**Naina Fathima K.**

# Abstract

One of the major environmental issues brought on by the growth of urbanisation and industrialisation is air pollution. PM2.5, PM10, CO2 and other air pollutants are among the primary ones. The effects of air pollution have a major negative influence on the ecosystem and the health of the global population. The fourth-highest risk factor for human health on the planet is air pollution. It is to blame for roughly 16% of all deaths globally. Monitoring and improving air quality is considered one of today's biggest challenges.

The project AirGuardian introduces an innovative system that utilizes a combination of Internet of Things (IoT) technology and machine learning algorithms to predict air quality and implement a proactive disease warning system. By deploying a network of IoT sensors strategically throughout targeted areas, the system continuously collects real-time data on various air pollutants, including particulate matter, nitrogen dioxide, sulphur dioxide, carbon dioxide, methane, carbon monoxide, and ozone. The system's core objectives are twofold: monitoring air quality and forecasting potential health risks associated with poor air quality. Leveraging historical data, the system offers insights into the likelihood of diseases linked to air pollution, facilitating proactive public health management strategies. The incorporation of Explainable AI ensures that the system's predictions are transparent and easily understandable to end-users, including healthcare professionals and the general public. This adaptable and scalable solution provides actionable insights for air quality prediction, applicable across diverse urban environments, ultimately enhancing public health and environmental sustainability.

# Contents

# List of Abbreviations

IoT - Internet of Things

LSTM - Long Short Term Memory

SHAP - SHapely Additive Explanations

# List of Figures

# Chapter 1

# Introduction

## 1.1 Background

Our project focuses on the growing concern about worsening air quality worldwide. Factors like increased industrialization, urbanization, and more vehicles on the road are pumping more pollutants into the air, affecting both the environment and people's health.

Existing air quality monitoring systems, while valuable, often face limitations in real-time accuracy and spatial coverage. To address these challenges, our project leverages advancements in IoT technology and machine learning models to create a comprehensive air quality monitoring and prediction system.Explainable AI is then used to explain the model.

This initiative not only contributes to the scientific understanding of air quality dynamics but also holds practical implications for public health. Through the identification and prediction of harmful gases, our project endeavors to forecast potential health risks and associated diseases, thereby facilitating proactive measures for mitigating the impact of air pollution on communities.

## 1.2 Problem Definition

Develop an integrated system for air quality analysis and monitoring combines machine learning models trained on historical air gas concentration data with real-time IoT data collection to identify and mitigate harmful air pollutants, utilizing explainable AI techniques for transparent insights into decision-making processes.

## 1.3 Scope and Motivation

The project focuses on understanding air quality using advanced technology and data analysis. Machine learning is utilized to predict air quality and conduct in-depth studies on greenhouse gases. Data is gathered from 2013 Air Quality data from Beijing and real-time IoT sensors to enhance understanding of air quality fluctuations. The aim extends beyond mere prediction to the identification of harmful gases, facilitating environmental and public health monitoring. The potential applications include urban planning and policy development in support of the environment.

The project is driven by the urgent need to tackle worsening air quality and its effects on health. Pollution levels are increasing globally, prompting the search for proactive solutions. Employing modern technology, the aim is to provide communities and policymakers with the necessary tools for informed decision-making regarding public health and environmental concerns. The project is dedicated to surpassing traditional monitoring methods to gain a deeper comprehension of how atmospheric conditions and harmful gases influence health. Ultimately, the aspiration is to foster positive change and promote collaborative efforts towards a cleaner, healthier future.

## 1.4 Objectives

- Compare existing machine learning models to develop a model for precise air quality prediction and analysis.

- Portray influence of individual factors contributing to air quality using Explainable AI.

- Collect real time data through IOT system using gas sensors.

- Implement disease warning system based on concentrations of harmful gases in the environment.

### 1.5 Assumption and Challenges

The project relies on several key assumptions.Firstly, the project depends on the reliability of sensors for accurate data collection. The assumption is that the models can effectively apply their learnings to new data. Timely data transmission is also crucial for ensuring real-time relevance. Lastly, there's an assumption of sufficient computing and storage resources for optimal system functionality.

However, the project encounters multiple challenges. Real-time IoT data collection faces issues like intermittent connectivity and sensor malfunctions, affecting overall accuracy. Balancing model accuracy with computational efficiency and budget constraints are practical challenges. Addressing these challenges is vital for the success of the proposed air quality monitoring system.

### 1.6 Societal / Industrial Relevance

Our project directly benefits society by improving air quality monitoring and prediction, which is crucial for public health. The accurate forecasts generated by our integrated system empower individuals, communities, and healthcare professionals to take timely actions in response to changing air quality conditions. This proactive approach helps mitigate the adverse health effects associated with air pollution, ranging from respiratory issues to broader public health concerns. In the industrial landscape, our project aligns with the growing emphasis on sustainable practices and environmental responsibility. By utilizing the dataset and real-time sensor inputs, optimized through machine learning, industries can gain precise insights into their environmental impact. This enables them to adopt targeted measures to reduce emissions and pollutants, meeting regulatory standards and contributing to a more sustainable future.

## 1.7 Organization of the Report

The report is organized into several key sections. The introduction provides a background,followed by problem definition, outline of the scope and motivation behind the project. Following this, the assumptions made throughout and the challenges faced are mentioned. A thorough literature survey is conducted, analyzing five seminar papers relevant to the subject matter which includes the base paper. The subsequent section involves the requirements, system overview and implementation,architecture,work schedule and results. Finally, the conclusion is made outlining the future scope and extensions.

# Chapter 2

# Literature Survey

## 2.1 Sensing Data Fusion for Enhanced Indoor Air Quality Monitoring[1]

In order to analyse the indoor air quality,we are using a different approach from the standar air quality index that is use for predicting outdoor air quality.An air quality index scale is generally used that is color coded for ease of understanding.This research put forward the importance of expanding this to include indoor surroundings as it has significant influence on human well being.

Apart from the contaminats measured in outdoor air quality,additional pollutants are also measured.The pollutants are chosen according to their influence on air quality taking into account how their concentrations are affected by ventilation and human activity.The improved IAQ assesment,provides a better understanding of air quality succh as homes,confined areas etc.

The process starts with reading data using waspmote sensors,a sensor network developed by the company Libellium.It takes data from CO2,NH3,H2S,CO,C7H8,C2H60 sensors during regular intervals.It also acknowledges O2,relative humidity and temperature.In order to make the sensor more reliable,calibration is done that guarantees alignment between sensor measurements and reference values.This calibrated device is used in smart buildings. The Extended Fractional Kalman Filter (EFKF) is used to process the raw data obtained from the building's IAQ sensor network. The EFKF serves as a filtering approach, increasing the accuracy of indoor air quality (IAQ) information obtained from the sensors. It was selected due to its effectiveness in managing the nonlinear and stochastic characteristics present in sensor data.

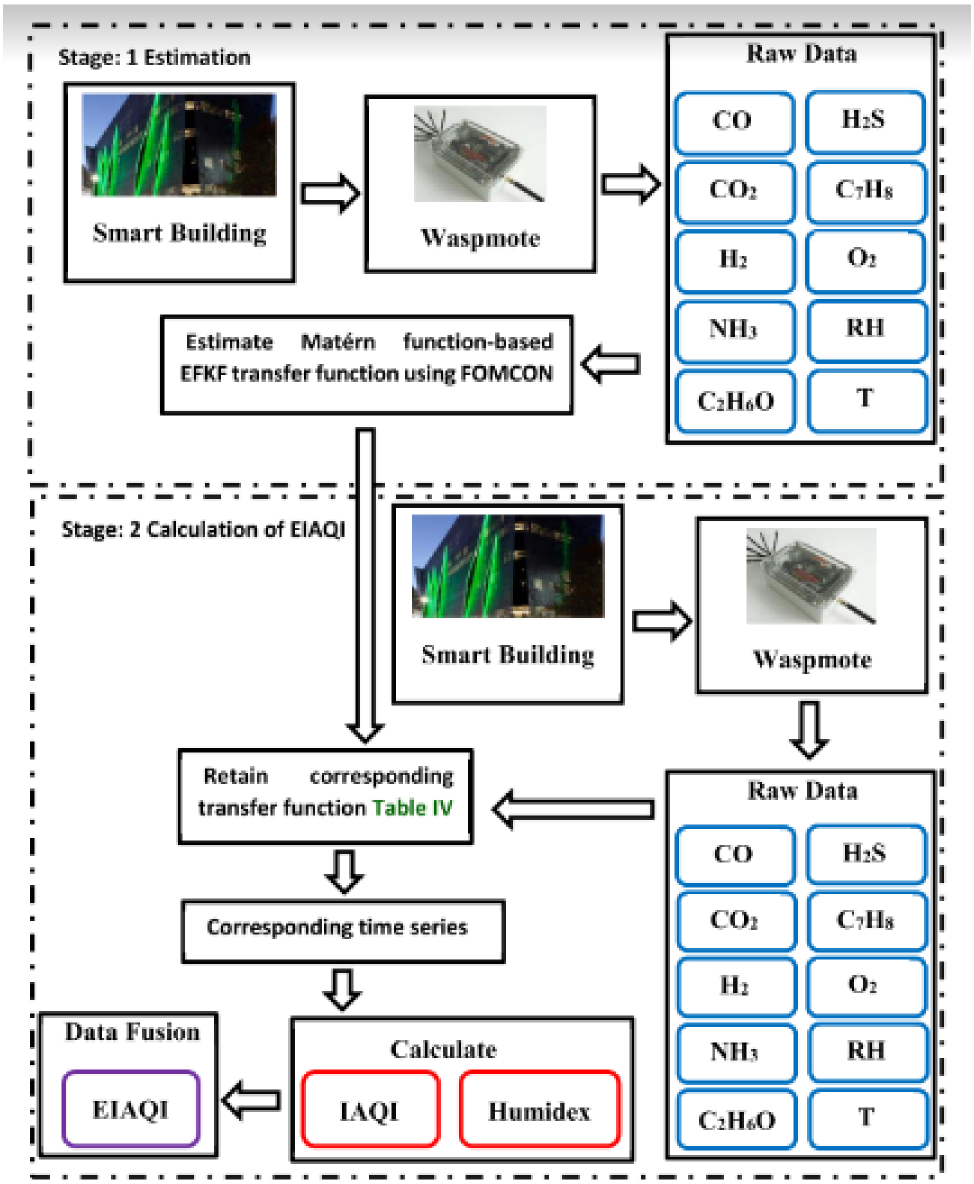



Figure 2.1: Data Fusion[1]

The Matérn function is used in the EFKF framework to model and calibrate sensor data while taking spatial correlations into consideration. This feature is essential for identifying complex relationships between different sensor data and for giving a comprehensive picture of the dynamics of indoor air quality.

Fractional Order Modeling and Control (FOMCON) is a toolbox that facilitates accurate modeling of systems. Fractional-order systems can be identified and their parameters estimated with the help of this toolbox, guaranteeing that the final transfer function appropriately captures the dynamics of the IAQ environment.

The calibration process determines the coefficients for the transfer function (F(s)) using data on pollutant levels gathered at the testbed building. This matches the behavior of indoor air pollutants, which are expressed as a ratio of polynomials, in the real world with the mathematical model.

Fractional Order Modelling and Control takes the pollutant level data as input and classifies the behaviour of indoor air pollutants.The correctness of the model is checked using mean square error. A linear interpolation formula is used for calculating air quality index.Another special interpolation formula is used for calculating oxygen.The humidex,which was developed by a Canadian meterologists measures how a person feels coolness or hotness.

The obtained indoor air quality and humidex is fused together to form enhances indoor air quality.This is done by giving weights to both and adding it together to obtain the final enhanced indoor air quality.The weights range from -2-3 and represents the importance of each factor.

This research concludes with a thorough technique for evaluating and forecasting indoor air quality. It tackles the intricacies of pollutant dynamics by expanding the traditional Air Quality Index (AQI), adding Humidex, and employing Extended Fractional Kalman Filtering (EFKF) for sensor data. Through practical applications of accurate indoor air quality measurement, the study highlights the significance of ongoing monitoring for building management.

### 2.2 Air Quality Forecasting Using the GRU Model Based on Multiple Sensors Nodes[2]

In this paper, a way of choosing neighbors that takes Mutual Information (MI) into account is put forward to increase the accuracy of air quality predictions in wireless sensor networks. MI is an effective means of measuring the information exchange between different sensor nodes. The method cleverly exploits MI to find key nodes closely related to the target node's data, making sure that this selection of neighboring points has high spatial correlations. This fine-tuned node selection process offers a number of advantages. Secondly, it only considers the information most relevant to nodes that strongly affect the comprehension of a given node's data. Further, MI permits a significant reduction in dimensionality. The analysis can focus on only dim nodes of high mutual information and thereby treat subsequent modeling tasks easier to handle.

Then, based on the selected nodes, a lightweight Gated Recurrent Unit (GRU) network is used for time-series prediction. For capturing the temporal dependencies in sequential data, this network architecture with two GRU layers connecting through a dropout to reduce overfitting and then finally flowing into linear connections is especially suited. The input to the GRU network is constructed by appending time-series data for chosen neighboring nodes with those of target node. Its network output is a set of predicted air quality parameters for the next moment. The optimization technique uses Adam optimizer and Mean Squared Error (MSE) Loss function, adjusting the learning rate dynamically to smooth convergence.In practice implementing this method in a wireless sensor network environment, nodes equipped with multiparameter sensing modules receive varying environmental parameters from end-devices. The nodes send their data to a sink node, which consists of the main control board and Raspberry Pi 3B+. The sink node executes the trained GRU model for real-time forecasting of indoor air quality. This holistic approach makes use of the spatio-temporal correlations in sensor data to produce more reliable and accurate predictions. Moreover, the methodology tackles problems including sensor failures and environmental disturbances, which make for a more robust predictive model. In summary, the combination of MI-guided neighbor node selection and GRU network training forms an overall framework for spatio-temporal prediction of air quality. MI is highly adaptable to changes in network conditions, and GRU networks can offer very efficient learning. Combined with the important fact that fluctuations in air quality will have a big impact on people's health when they are indoors all day long, this approach is particularly suited if it leads to real-world applications.

## 2.3 PM2.5 Prediction Using Genetic Algorithm-Based Feature Selection and Encoder-Decoder Model[3]

The proposed prediction model which is designed to predict PM2.5 combines an encoder-decoder (E-D) architecture with feature selection via the use of Genetic Algorithms (GA). The GA finds the highest-yielding feature combinations by expressing features as binary strings and iterative fine-tuning of the population through crossover and mutation procedures. Mean Absolute Error (MAE) is a quantitative measure of fitness, which leads the process of choosing persons for future generations. Over iterations, this heuristic approach improves the entire population's fitness; during crossing it ensures that good characteristics are never lost.

The E-D model 'LSTM units carry out the encoding and processing of GA selection features to simulate this sequence. From historical features, the encoder defines an encoder state. This state is used by the LSTM-based decoder to predict PM2.5 levels for next time steps. The LSTM units solve the vanishing gradient problem, and it also makes easier to find long-term dependencies in time series data. Gated by features, the E-D model uses a dense layer to produce final PM2.5 forecasts.

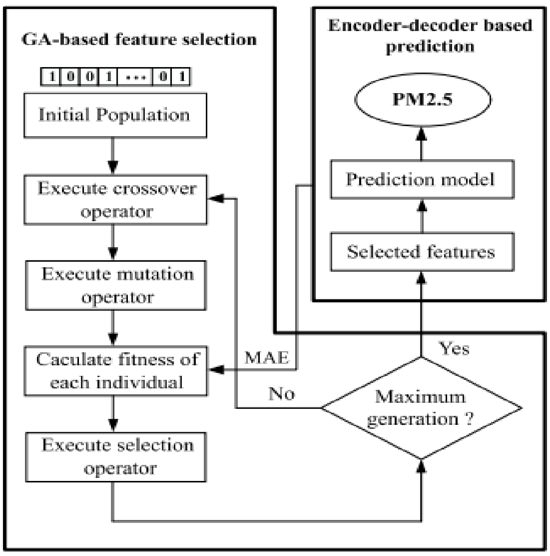



Figure 2.2: Workflow of E-D Model[3]

In this technique, the GA iteratively evolves feature combinations that increase fitness. With the selected features, we train E-D model; PM2.5 predictions keep improving. This method can also combine LSTM features in capturing temporal relationships with GA's facility for feature selection, thus providing a powerful framework to obtain accurate PM 2.5 forecasting results.

Using many benchmarks, the accuracy of a proposed ED-LSTM model with feature selection based on Genetic Algorithms (GA) is first assessed. Second, using datasets from Taiwan and Hanoi this GA-based feature selection is compared with other methods such as XGBoost and Pearson's correlation. Compared to them the GA approach clearly wins out in reducing Mean Absolute Error (MAE). Second, ED-LSTM significantly outperforms current models of AE-BiLSTM and AC- LSTM in terms of prediction accuracy; improvements range from 14.7 percent to 53.7 percentage points (in MAE). The superiority of the model over ST-DNN, a cutting-edge model for this task, shows that it is effective across various time step forecasts. The E-D LSTM architecture in conjunction with GA feature selection is a reliable means of predicting PM2.5.

## 2.4 Air Quality Prediction Based on Integrated Dual LSTM Model[4]

The method for air quality prediction is an integrated dual LSTM (Long Short-Term Memory) model. The methodology simultaneously exploits the strengths of single- and multi-factor models in order to improve air quality predictions. The results are combined using an XGBoosting tree.

The process involves several steps. Then a single-factor prediction model is built using Seq2Seq technology. In this model, every element of air quality is a time series and its values can be predicted independently. LSTM unit-based encoding and decoding forms the basis of Seq2Seq model. One can infer the temporal patterns and characteristics of each component by training the model on historical data.

Then an attention-based multi-factor prediction model is constructed using LSTM. There are many factors which affect air quality, including weather and data from nearby stations. This model takes all these variables into account. With an attention mechanism, the model can focus on those key variables that are most important for prediction. The multi-factor model employs the Seq2Seq structure with LSTM units for encoding and decoding as well. By integrating a number of parameters, the model can represent time and space relationships found in air quality data.

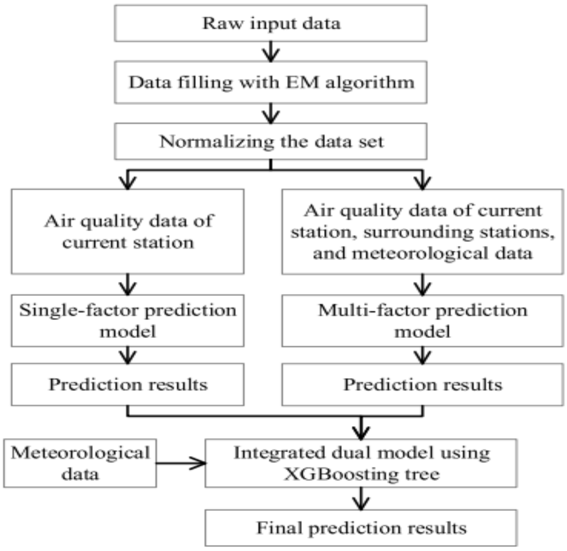



Figure 2.3: Workflow of Integrated Dual LSTM Model[4]

An XGBoosting tree is used to combine the output from the single-factor and multi-factor models. XGBoosting is a gradient boosting strategy that builds a powerful ensemble model by combining several weak models. Here, the output of the multi-factor spatial model and the single-factor time model are fed into the XGBoosting tree. To get the optimal expected value, the tree determines the weight of each leaf node and adds up its predicted value. The accuracy of the predictions overall is increased by this integration stage.

The methodology additionally tackles the problem of absent data in the dataset on air quality. This assignment is not suited for traditional data filling techniques like mean filling or deletion. Rather, to fill in the gaps, the EM (Expectation-Maximization) method is employed. Based on insufficient data, the EM algorithm fills in the gaps in the data to estimate a probability model's parameters. It contributes to keeping the final data's distribution probability close to that of the original data.

Five assessment indicators are used to assess the efficacy of the suggested methodology: R-square, IA (Index of Agreement), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), and RMSE (Root Mean Square Error). The integrated dual LSTM model performs better than other models in terms of error, according to the results.

The authors intend to increase the integrated dual LSTM model's use in future work in order to raise the precision of diverse data predictions. Even though outliers are uncommon in their model, they nonetheless seek to address the problem of outlier values in the prediction results. In general, the suggested approach exhibits potential for enhancing the precision of air quality forecasts and may be used on diverse datasets possessing numerous attributes.

### 2.5 Spatiotemporal prediction of air quality based on LSTM neural network[5]

The study's methodology revolves around the development and evaluation of a Multivariate Multistep Long Short-Term Memory (MMSL) prediction model for air quality concentrations, with a specific focus on Beijing's urban environment. The pressing need for precise air quality prediction is underscored, given the escalating environmental pressure stemming from rapid societal development, particularly in the form of air pollution, which has significant implications for public health, notably cardiovascular disease.

The first step in building the MMSL model is gathering data from 35 monitoring stations that span a range of environmental assessment points. The methodology stresses that the problem of missing data is very prevalent, and one important step towards solving it should be cleaning up available data. A number of different methods alternative for filling in the empty data; interpolation turns out to be most effective, superior even over fixed values (constant), means and modes. Furthermore, factors such as data normalisation and scaling to a [0, 1] range have the effect of raising both the model's accuracy and its rate of convergence.

Spatiotemporal analysis is an important element in the methodology, and rests on

Tobler's First Law of Geography. which states that things closer together tend to be correlated more than things further apart. Based on latitude and longitude, the Haversine formula is used to calculate distances between monitoring stations. The linear correlation coefficient is then used to measure the strength of the relationship between two variables. This research provides a considerable degree of understanding about the spatial relationships between monitoring stations, which is an important prerequisite for high-accuracy air quality forecasting.

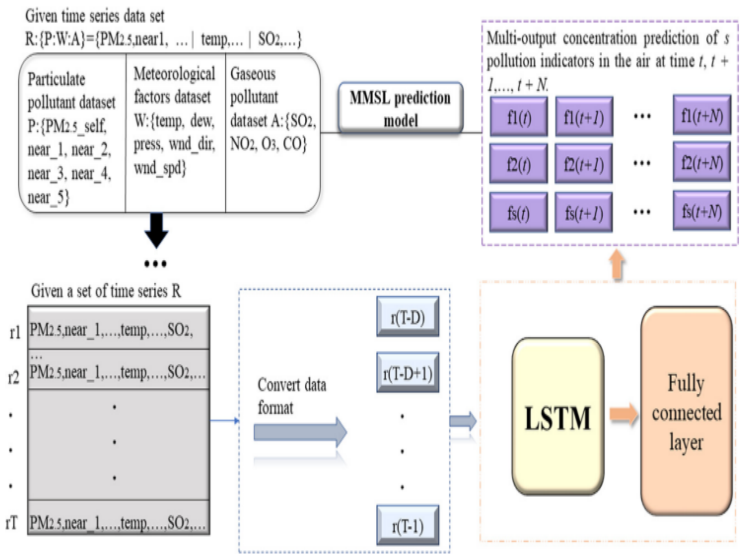



Figure 2.4: Workflow of LSTM Model[5]

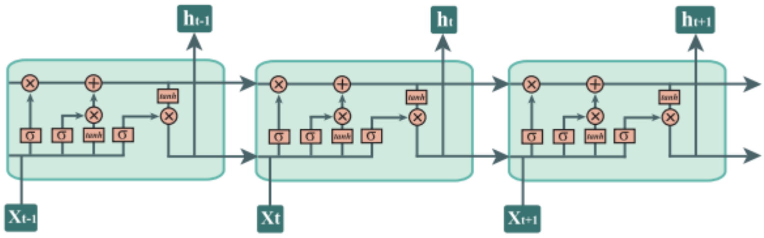
The MMSL is one of the main ingredients in this methodology. It endeavors to include past pollutant concentration,meteorological factors and interdependences between pollutants as well in its calculations. The time series prediction problem is precisely stated, in which the data sets on pollutant concentrations and meteorological factors are disjoint, while those for gaseous pollutants overlap with both of these. Each set has 15 features. The authors generate target prediction time series, describing the relationship between observed and predicted values. In order to predict future pollutant levels from historical data, the method uses lagged time series.Predictive accuracy is measured with Mean Absolute Error (MAE), Root Mean Squared Error (RMSE). and the Decision Coefficient or R-squared (R2) as key metrics To evaluate how well a model performs in making projections of this type for

Finally, this study's methodology represents a systematic and rigorous way to forecast air quality. It integrates spatiotemporal analysis, data preprocessing and an advanced deep learning model. The detailed comprehensiveness or advanced methodologies applied in the MMSL prediction model mean that it is robust and effective.

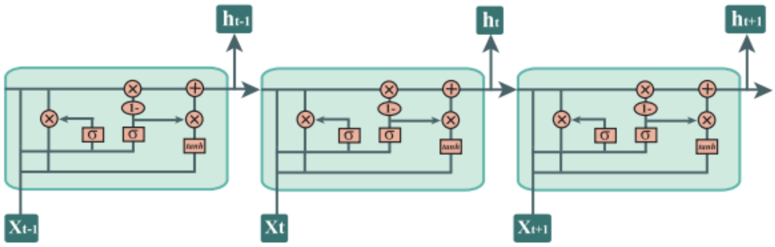
## 2.6 Revealing Influence of Meteorological Conditions on Air Quality Prediction Using Explainable Deep Learning[6]

The paper uses explainable deep learning to reveal the impact of meteorological conditions on air quality prediction. The prediction accuracy is higher when meteorological conditions are combined with other air pollutants. The investigated results can help improve the accuracy of air quality predictions.

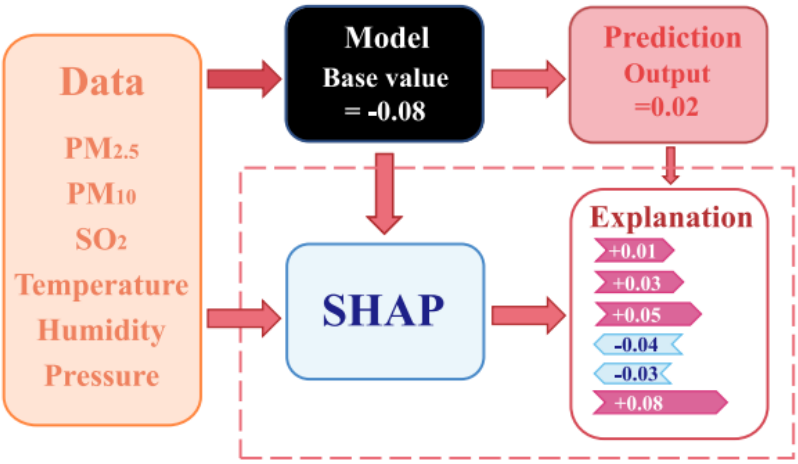
LSTM and GRU models are commonly used to predict air quality. To understand how these models work, the SHAP method is applied for explainability. LSTM is a type of Recurrent Neural Network (RNN) that excels at remembering long-term patterns. On the other hand, GRU, or Gated Recurrent Unit, is another variant of the LSTM model. SHAP, a game-theoretic approach, helps us comprehend the models' outputs by revealing the impact of different features in each prediction. It indicates whether a feature has a positive or negative influence on the model's output. It connects optimal credit allocation with local explanations by the classical Shapley values from game theory and the related extensions. Single-sample analysis is performed to analyze the distribution of SHAP values for individual features and individual samples. Analysis for all features is then performed to analyze the importance distribution of all features in the air quality prediction to have a more intuitive understanding of the influence of meteorological conditions on air quality prediction.




(a) LSTM model.




(b) GRU model.




(c) SHAP model.

Figure 2.5: SHAP Model Diagrams[6]

The results show that (1) in both the LSTM and GRU models, the prediction accuracy is not improved by considering only meteorological conditions. However, when considering other air pollutants, the prediction accuracy is improved, and when combining meteorological conditions with other air pollutants, the prediction accuracy is even higher. (2) Whether only considering meteorological conditions or combining meteorological conditions and other air pollutants for PM2.5 prediction, in both the LSTM and GRU models, the meteorological conditions have a high contribution to air quality prediction.

## 2.7 Air Quality Monitoring and Disease Prediction Using IoT and Machine Learning[7]

This project focuses on the development of an IoT-based air quality prediction system using machine learning algorithms. The system aims to monitor air pollutants and predict their impact on human health, including the possibility of diseases such as asthma, lung cancer, and ventricular hypertrophy. The proposed system utilizes high-end sensors to measure various air pollutants and employs multi-label classification with Random Forest and XG Boost algorithms for predictive modeling. The results obtained from the system include the Air Quality Index (AQI) and the diseases caused by observed data. Preventive suggestions are also provided to mitigate the effects of pollutants.

Air pollution is a significant concern, with pollutants such as particulate matter, carbon monoxide, sulfur dioxide, and nitrogen oxide posing serious threats to human health. The project aims to address this issue by developing an IoT prototype equipped with high-end sensors to monitor air pollutants and predict their impact on human health. The system utilizes machine learning algorithms for predictive modeling and provides preventive suggestions to reduce the effects of pollutants.

The project builds upon previous research in the field of air quality monitoring and disease prediction. Various methods and technologies have been proposed, including the use of data mining algorithms, particulate matter monitors, sky luminance estimation, and color channels for detecting air pollution. Additionally, previous works have focused on air quality monitoring using IoT devices and cloud environments, as well as the analysis of air data for predicting health risks.

The proposed system involves the development of an IoT-based air quality prediction system that utilizes high-end sensors to measure air pollutants. The system employs machine learning algorithms, such as Random Forest and XG Boost, for predictive modeling and aims to predict the possibilities of diseases and calculate the Air Quality Index (AQI). The system also provides preventive suggestions to mitigate the effects of pollutants on human health.

The results obtained from the system include the Air Quality Index (AQI) and the diseases caused by observed data. The system utilizes multi-label classification models to predict the impact of air pollutants on human health, including the possibility of diseases such as asthma, lung cancer, and ventricular hypertrophy. Visual representations of the variation of AQI values in different locations and the diseases caused in different regions are also presented.

The project concludes that the proposed IoT-based air quality prediction system has the potential to effectively monitor air pollutants and predict their impact on human health. The system's use of machine learning algorithms and high-end sensors enables the accurate prediction of diseases and the calculation of the Air Quality Index (AQI). The project emphasizes the critical importance of maintaining stable pollutant levels and mitigating the effects of industrial pollutants on human health.

# Chapter 3

# Requirements

## 3.1 Hardware and Software Requirements

**Hardware Requirements:**

- Arduino Uno:

  The Arduino Uno is a widely used open-source microcontroller board, featuring an ATmega328P microcontroller, equipped with digital and analog input/output pins, making it versatile for a broad range of projects in the fields of electronics and programming.

- MQ131 Ozone Sensor:

  The MQ131 is a gas sensor specifically designed for detecting ozone (O3) levels in the air, providing an analog output voltage that varies based on the concentration of ozone gas.

- DHT11 Sensor:

  The DHT11 sensor is used to measure temperature and humidity in various applications

- MQ135 Sensor:

  The MQ135 sensor is utilized to detect a range of gases including ammonia, benzene, alcohol, smoke, and carbon dioxide (CO2).

- MQ7 Sensor:

  The MQ7 sensor is employed to detect carbon monoxide (CO) levels in the atmosphere, making it valuable for applications involving gas leakage detection, indoor air quality monitoring, and safety systems where CO detection is critical.

- ESP8266 :

  The ESP8266 is a popular Wi-Fi module and microcontroller commonly used for IoT and home automation projects.

- Connecting Wires:

  Various connecting wires to establish electrical connections between Arduino and sensors.

**Software Requirements:**

- Visual Studio Code (VS Code):

  An integrated development environment (IDE) for coding, debugging, and version control. Supports Python development and extensions for IoT development.

- Python:

  Python serves as the primary programming language for developing machine learning models, data processing, and server-side scripting. Recommended version: Python 3.x. Machine Learning Libraries:

  Machine learning libraries such as Scikit-learn is used for building predictive models based on air quality data.

- Jupyter Notebook :

  Jupyter Notebook is a tool for interactive computing, widely used for coding, data analysis, and sharing documents containing code, visualizations, and text.

- Arduino IDE :

  This is the primary software for programming Arduino boards.

# Chapter 4

# System Architecture

### 4.1 System Overview

AirGuardian is an innovative project combining IoT and AI to revolutionize air quality monitoring. Using IoT sensors, the system collects real-time environmental data and stores it in the cloud. The Air Quality Prediction module, powered by LSTM, forecasts individual gas concentration levels. The Explainable AI module ensures transparency in the prediction process. Lastly, the Disease Warning System alerts users of potential health risks associated with the predicted air quality, linking gas concentrations to possible health concerns. AirGuardian offers a comprehensive approach to understanding and addressing air quality challenges for public well-being.

The IoT module is designed with a meticulous approach to effectively monitor gases, such as carbon dioxide, methane, sulphur dioxide, nitrogen dioxide, ozone as well as ambient temperature, humidity, dust particle levels. At the core of this system is the utilization of an Arduino board, which functions as the central control unit. Equipped with sensors tailored for gas, temperature, and humidity measurements, the Arduino board serves as the backbone of the entire circuit, forming the basis for data collection and control.For seamless real-time data transmission, the Arduino board is integrated with a WiFi module. This module plays a crucial role by providing wireless connectivity, enabling the swift transfer of data from the onboard sensors to the cloud. Beyond facilitating communication, the WiFi module supports continuous data streaming, ensuring that the system captures and transmits dynamic changes in the monitored environmental parameters in real-time.

The collected data is securely stored in a cloud-based storage solution. This approach offers several advantages, including scalability, accessibility, and redundancy. By leveraging cloud storage, the system ensures that the data is not only preserved efficiently but is also readily available for subsequent analysis.

The Long Short-Term Memory (LSTM) prediction model is designed for time series forecasting, specifically aimed at predicting air quality measurements. The dataset is inherently temporal, with each row representing a distinct time point, encompassing details such as year, month, day, and hour. The key pollutants, including PM2.5, PM10, NO2, SO2 and ozone form the basis of the model input. The initial stages involve meticulous data preprocessing steps, including the removal of irrelevant columns and the handling of null values. Subsequently, the dataset is split into training and testing sets, crucial for training and evaluating the model's performance. The pollutants PM2.5, PM10, NO2, SO2 and ozone, the target variables, undergo normalization to ensure a consistent range for all input features. The concept of time series prediction is manifested in the sequence splitting process. The time series data is converted into input-output pairs. Each input sequence represents a historical segment, and the corresponding output is the target variable to be predicted. The number of time steps in each sequence plays a pivotal role in determining the amount of historical information the model considers. The LSTM model architecture, a form of recurrent neural network (RNN), is specifically tailored for processing sequential and time-dependent data. With a single LSTM layer containing 50 units and a 'relu' activation function, the model is well-equipped to capture long-term dependencies and temporal patterns within the time series data. The choice of Mean Squared Error (MSE) as the loss function is fitting for regression tasks, aligning with the goal of minimizing the squared differences between predicted and true values. During the training phase, the LSTM model learns from historical sequences to understand the temporal patterns inherent in the training data. The model's goal is to generalize this understanding to make accurate predictions for unseen future data.

The data obtained from gas concentration predictions made using LSTM can be analyzed to discuss the model's explainability, employing SHAP (SHapely Additive Explanations) for a comprehensive understanding. SHAP provides a clear and interpretable breakdown of the significance of each input variable, including PM2.5, PM10, NO2, SO2 and ozone. The Shapley values, derived through a cooperative game theory framework, quantify the average contribution of each feature across all possible permutations, offering insights into the relative importance of factors influencing air quality fluctuations. The computation of Shapley values involves evaluating the model's prediction for all possible combinations of features, comparing the predictions with and without each feature's inclusion. The Shapley value for a specific feature is the average difference in predictions across all possible permutations, reflecting its marginal contribution to the overall prediction. In this case, by summing the Shapley values, a holistic perspective on the factors affecting the predicted gas concentrations is gained. This comprehensive understanding allows us to prioritize and assess the relative importance of each environmental parameter, ultimately aiding in interpreting their individual impacts on the overall prediction.

## 4.2 Architectural Design

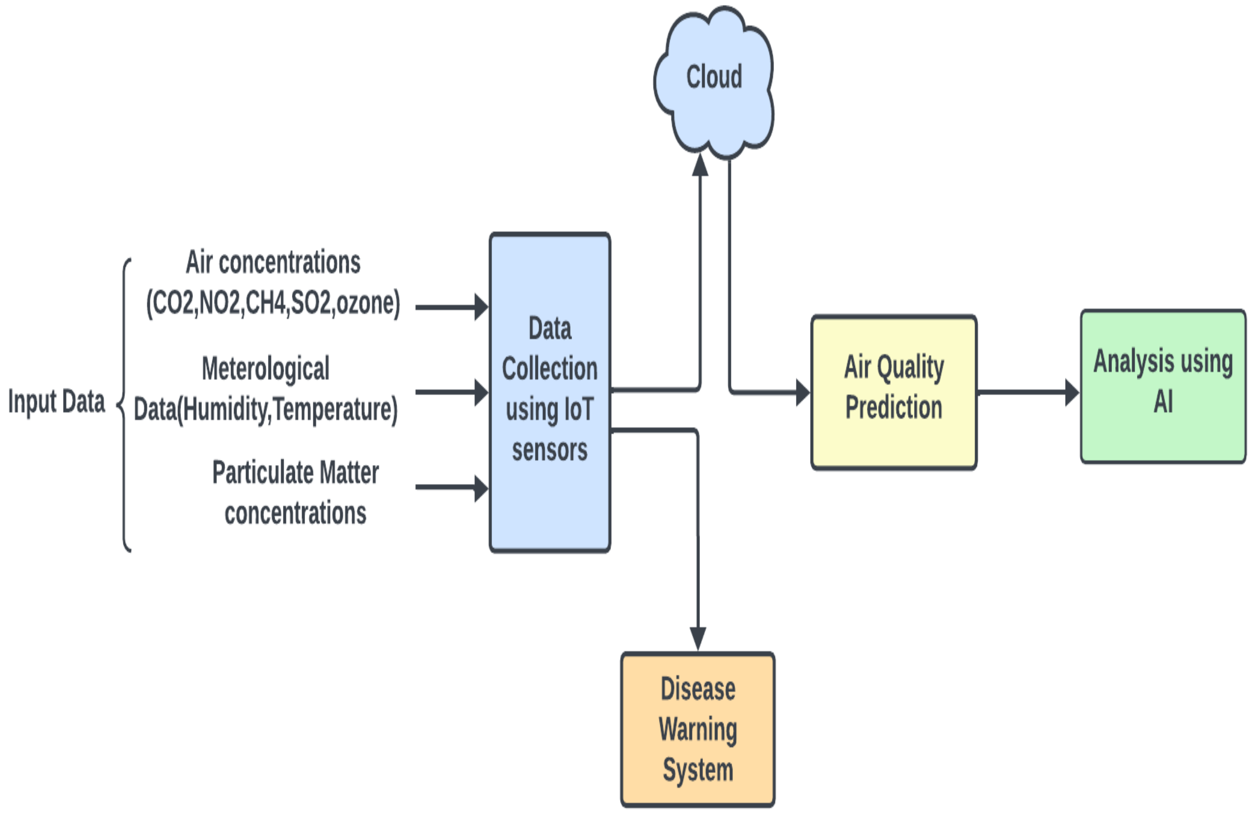



Figure 4.1: Architecture Diagram

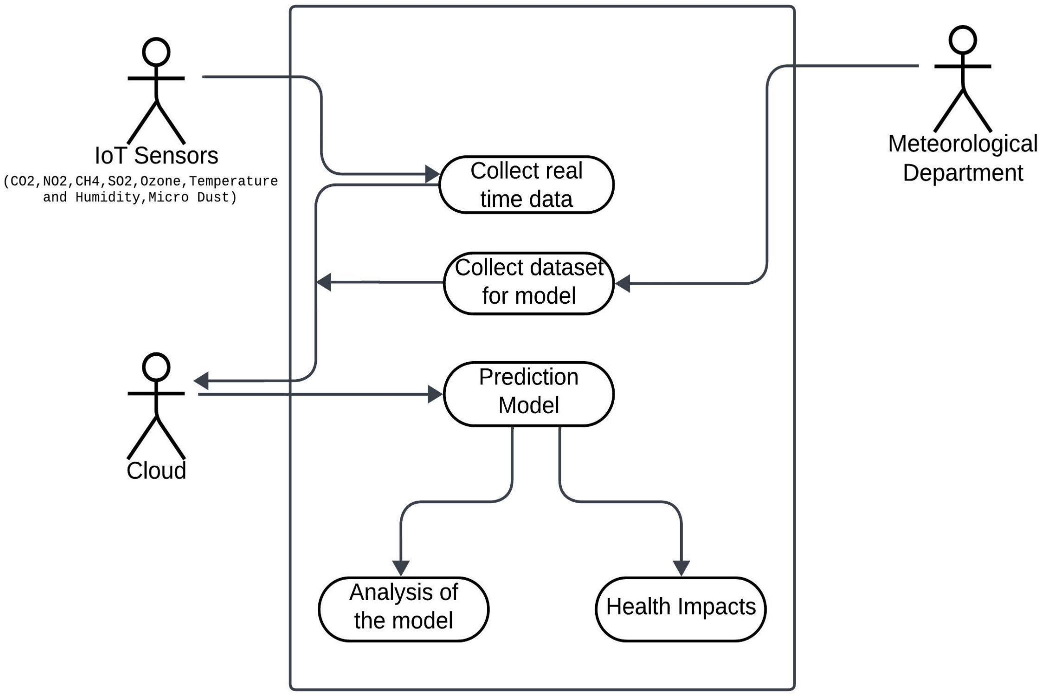



Figure 4.2: Use Case Diagram

### 4.3 Module Division

The project unfolds across four interconnected modules. The first module centers on IoT data acquisition, where real-time information is gathered from sensors to provide a dynamic input stream. The second module focuses on predicting Air Quality through advanced machine learning techniques. The third module delves into a disease warning system, utilizing analytics to anticipate health outcomes based on the collected data. The final module scrutinizes model decisions through Explainable AI, employing methods like SHAP to elucidate the impact of individual features.


Figure 4.3: Modulewise Diagram

### 4.4 Work Breakdown and Responsibilities


Figure 4.4: Work Breakdown and Responsibilities

## 4.5 Work Schedule - Gantt Chart

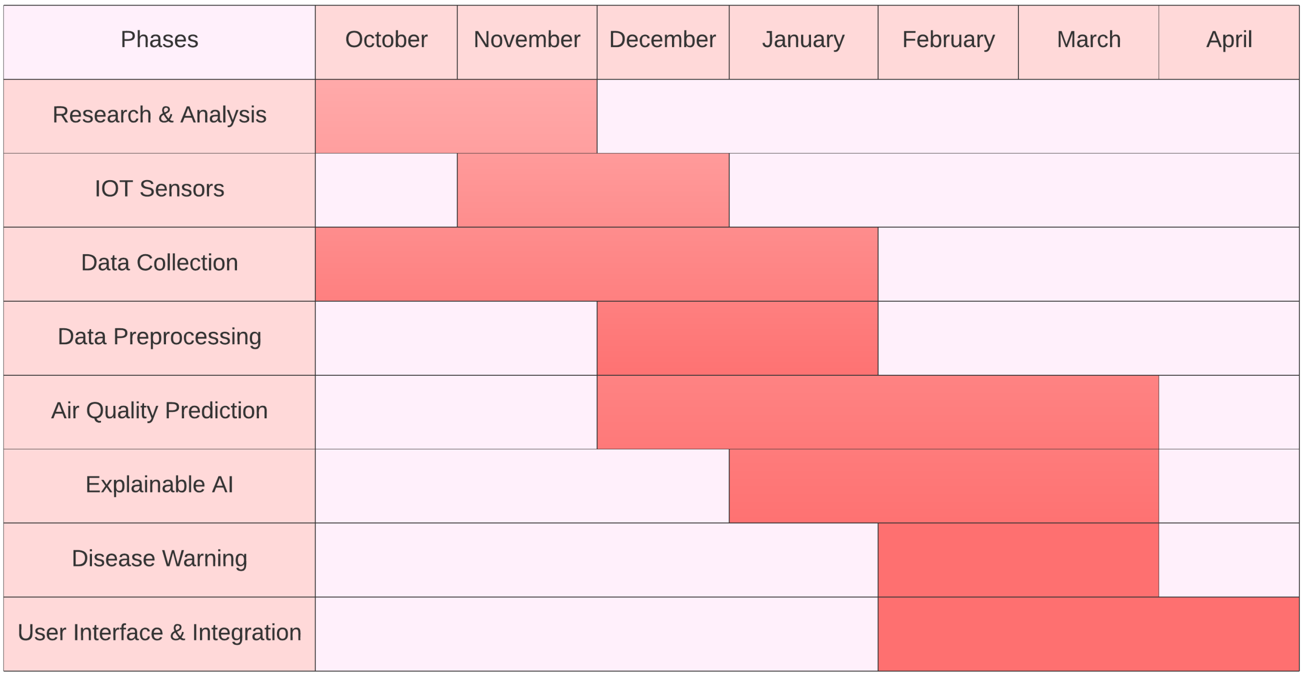



Figure 4.5: Gantt Chart

# Chapter 5

# System Implementation

## 5.1 Dataset Identified

The air-quality dataset is from the Beijing Municipal Environmental Monitoring Center. The meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. The time period is from March 1st, 2013 to February 28th, 2017.The columns in the dataset includes year,month, day, hour, PM2.5, PM10, SO2, NO2, CO, O3, with 34,043 rows of data.

## 5.2 Proposed Methodology/Algorithms

The project comprises four interlinked modules, each serving a distinct function. The initial module revolves around acquiring real-time data from IoT sensors, generating a continuous input stream. The second module is dedicated to predicting air quality using sophisticated machine learning methods. In the third module, a disease warning system is implemented, employing analytics to forecast potential health outcomes based on the gathered data. Lastly, the fourth module examines model decisions using Explainable AI techniques, such as SHAP, to clarify the influence of specific features.

### 5.2.1 Real-time data collection using IoT

The system implementation begins with the configuration of an Arduino board, which serves as the central control unit for gathering environmental data. Through integration with sensors like the DHT11 for temperature and humidity, MQ135 for CO2, methane, and NO2, MQ7 for CO and SO2, and MQ131 for ozone, the Arduino board collects real-time data on various air quality parameters. Additionally, softwareSerial communication is established with an ESP8266 WiFi module, enabling wireless transmission of the collected data. This setup forms the backbone of the IoT module, facilitating continuous monitoring of environmental conditions.

Once the hardware setup is complete, the system proceeds to configure the ESP8266 module for WiFi connectivity. Through a series of AT commands, the module is instructed to connect to the designated WiFi network using provided credentials. This step ensures that the Arduino board can seamlessly transmit data over the internet to the chosen cloud-based storage solution.

With connectivity established, the system enters a loop where sensor readings are obtained and formatted into a GET request. Each sensor reading,is assigned to a specific field in the GET request. This request, containing the API key for authentication, is then sent to the cloud server via the ESP8266 module.

To ensure reliable data transmission, the system employs error-checking mechanisms and waits for acknowledgment responses from the ESP8266 module after each AT command. This iterative process continues indefinitely, allowing for continuous monitoring and transmission of real-time environmental data to the cloud-based storage solution.

The integration with ThingSpeak provides a robust platform for securely storing and analyzing the collected data. ThingSpeak's visualization tools allow for easy monitoring of air quality parameters.

### 5.2.2 Disease Warning System

The disease warning system compares gas concentration levels with National Ambient Air Quality Standards (NAAQS) set by the Central Pollution Control Board (CPCB). By evaluating these values, the system determines the risk level, classifying it as high or low, which may be linked to specific diseases. This system plays a vital role in safeguarding public health by providing real-time alerts on potential health risks associated with air quality.

### 5.2.3 Air Quality Prediction

The Long Short-Term Memory (LSTM) prediction model is designed for time series forecasting, specifically aimed at predicting air quality measurements. The dataset is inherently temporal, with each row representing a distinct time point, encompassing details such as year, month, day, and hour. The key pollutants, including PM2.5, PM10, NO2, SO2, CO and ozone form the basis of the model input. The initial stages involve meticulous data preprocessing steps, including the removal of irrelevant columns and the handling of null values. Subsequently, the dataset is split into training and testing sets, crucial for training and evaluating the model's performance.

The pollutants PM2.5, PM10, NO2, SO2, CO and ozone, the target variables, undergo normalization to ensure a consistent range for all input features. The concept of time series prediction is manifested in the sequence splitting process. The time series data is converted into input-output pairs. Each input sequence represents a historical segment, and the corresponding output is the target variable to be predicted.The 3 number of time steps in each sequence plays a pivotal role in determining the amount of historical information the model considers. The LSTM model architecture, a form of recurrent neural network (RNN), is specifically tailored for processing sequential and time-dependent data. With a single LSTM layer containing 50 units and a 'relu' activation function, the model is well-equipped to capture long-term dependencies and temporal patterns within the time series data. The choice of Mean Squared Error (MSE) as the loss function is fitting for regression tasks, aligning with the goal of minimizing the squared differences between predicted and true values. During the training phase, the LSTM model learns from historical sequences to understand the temporal patterns inherent in the training data. The model's goal is to generalize this understanding to make accurate predictions for unseen future data. The model is then stored as a pickle file for later use as a predictive model.

### 5.2.4 Explainable AI

Explainable artificial intelligence (XAI) allows human users to comprehend and trust the results and output created by machine learning algorithms. It is used to describe an AI model, its expected impact and potential biases. It helps characterize model accuracy, fairness, transparency and outcomes in AI-powered decision making. SHapley Additive exPlanations(SHAP) is one such method of explainable AI. It is based on Shapley values, which use game theory to assign credit for a model's prediction to each feature or feature value. The way SHAP works is to decompose the output of a model by the sums of the impact of each feature. SHAP calculates a value that represents the contribution of each feature to the model outcome. These values can be used to understand the importance of each feature and to explain the result of the model to a human. The ML model is fed to the SHAP explainer to provide an analysis of the models outputs and how each feature influences the overall quality of the air. The SHAP explainer object is first set up by providing the model, then calculating SHAP value using a testing set.

The outputs from the explainer can be visually portrayed using a variety of graphs such as the summary plot, dependence plot, force plot, waterfall plot etc. The summary plot shows the feature importance of each feature in the model. A dependence plot is a type of scatter plot that displays how a model's predictions are affected by a specific feature. The force plot offers an in-depth perspective of SHAP values for individual instances. The Waterfall Plot is a useful visualization tool that displays the additive contributions of features to a model's prediction for a specific instance. The graphs obtained from the model indicate the strongest influence on air quality by the features PM2.5 and CO.

## 5.3 User Interface Design

The Website is designed to provide comprehensive insights into air quality conditions, utilizing real-time data collection from IoT sensors. This web-based platform comprises three primary pages: Home, Prediction, and Graphs, each serving distinct purposes to enhance user understanding and decision-making regarding air quality management.

### 5.3.1 Home Page

The Home page serves as the central hub of the website, offering real-time updates on gas concentrations derived from IoT sensors deployed across targeted locations. Additionally, the Home page hosts a disease warning system, which alerts users about potential health risks associated with prevailing air quality conditions.

### 5.3.2 Prediction Page

The Prediction page leverages the machine learning model to forecast gas concentration levels. This predictive functionality enhances the platform's utility by providing actionable insights for policymakers, environmental agencies, and individuals alike.

### 5.3.3 Graphs Page

The Graphs page offers visual representations of feature influences on air quality, utilizing SHAP explainable AI graphs for comprehensive analysis. By illustrating the correlations between various factors and air quality parameters, this page enhances user understanding of the underlying dynamics driving environmental conditions.

## 5.4 Conclusion

Our system develop a real-time air quality monitoring system. The dataset from the Beijing Municipal Environmental Monitoring Center is used, along with meteorological data for accurate analysis. The methodology includes four modules: real-time data collection using IoT sensors, air quality prediction using LSTM models, a disease warning system, and Explainable AI for feature analysis. The user interface design consists of three web pages - Home, Prediction, and Graphs - to provide insights into air quality conditions. This system aims to safeguard public health by offering real-time alerts and predictions, contributing to better air quality management.

# Chapter 6

# Results and Discussions

The system for Air Quality monitoring and prediction was built and tested.We checked its working using different datasets.The results and analysis are shown here.

## 6.1 Overview

In this section, we present the findings from our comprehensive air quality monitoring system, which tracks multiple gases and utilizes predictive modeling to estimate their concentrations. Furthermore, we explore the implications of these results, particularly in relation to public health, by comparing the obtained values with predetermined threshold levels to initiate a disease warning system.

## 6.2 Testing

### 6.2.1 Home Page-Website

The home page of the website comprises the real time data collected by the IOT setup as well as the disease warning system.

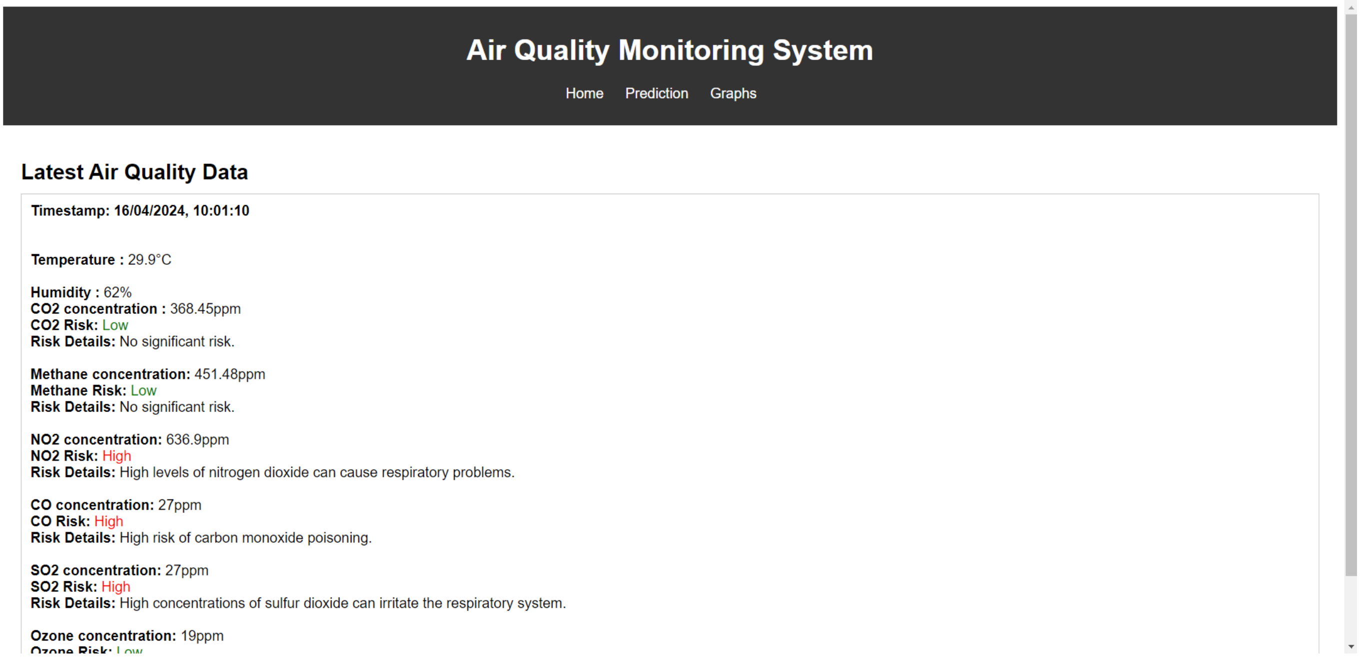



Figure 6.1: Website Home Page

Figure 6.1 displays the home page of the website showing the real time air concentrations data combined with the disease warning system.

### 6.2.2 Prediction Page-Website

The prediction page comprises the machine learning model which forecasts the gas concentration levels of the various features. Predictions of CO, O3 and SO2 are shown below.

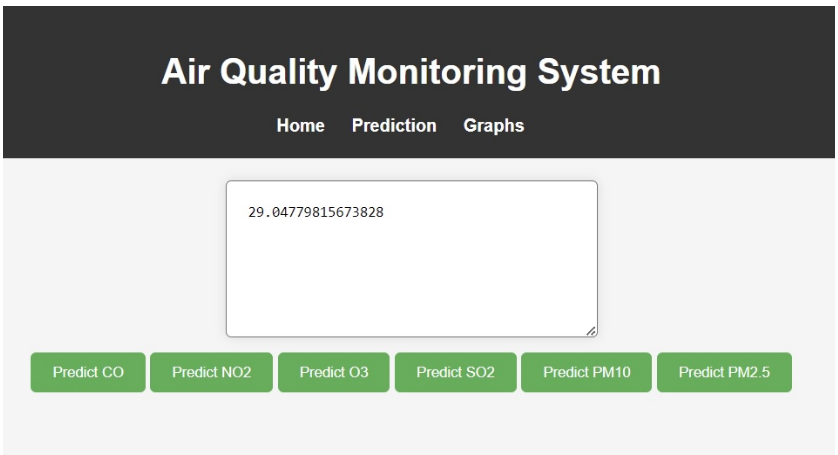



Figure 6.2: Website Prediction page-1

Figure 6.2 displays the prediction page of the website with the prediction of CO concentration.

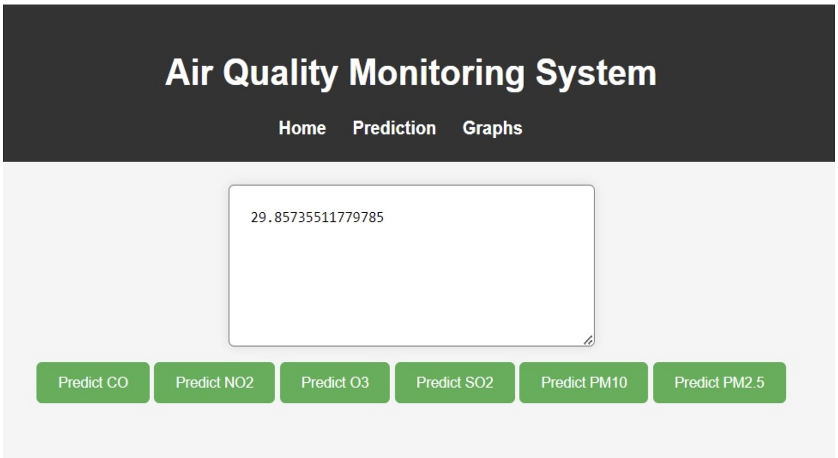



Figure 6.3: Website Prediction page-2

Figure 6.3 displays the prediction page of the website with the prediction of O3 concentration.

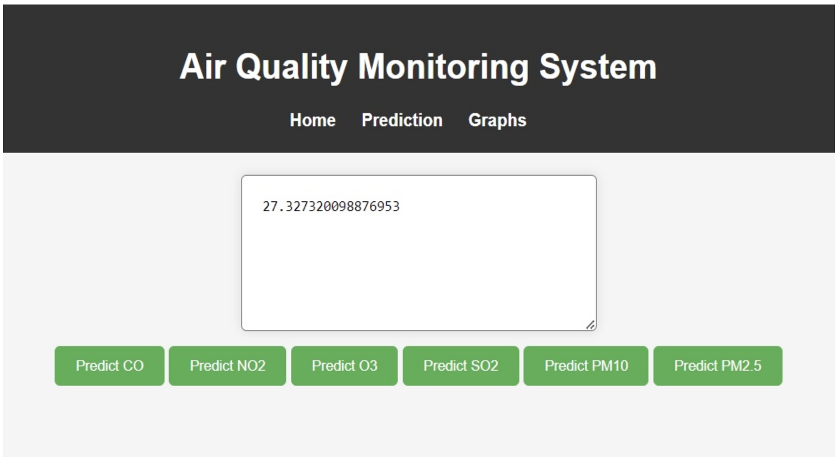



Figure 6.4: Website Prediction page-3

Figure 6.4 displays the prediction page of the website with the prediction of SO2 concentration.

### 6.2.3 Graphs Page-Website

The graphs page comprises of graphical representations of the influence of each feature to the air quality.

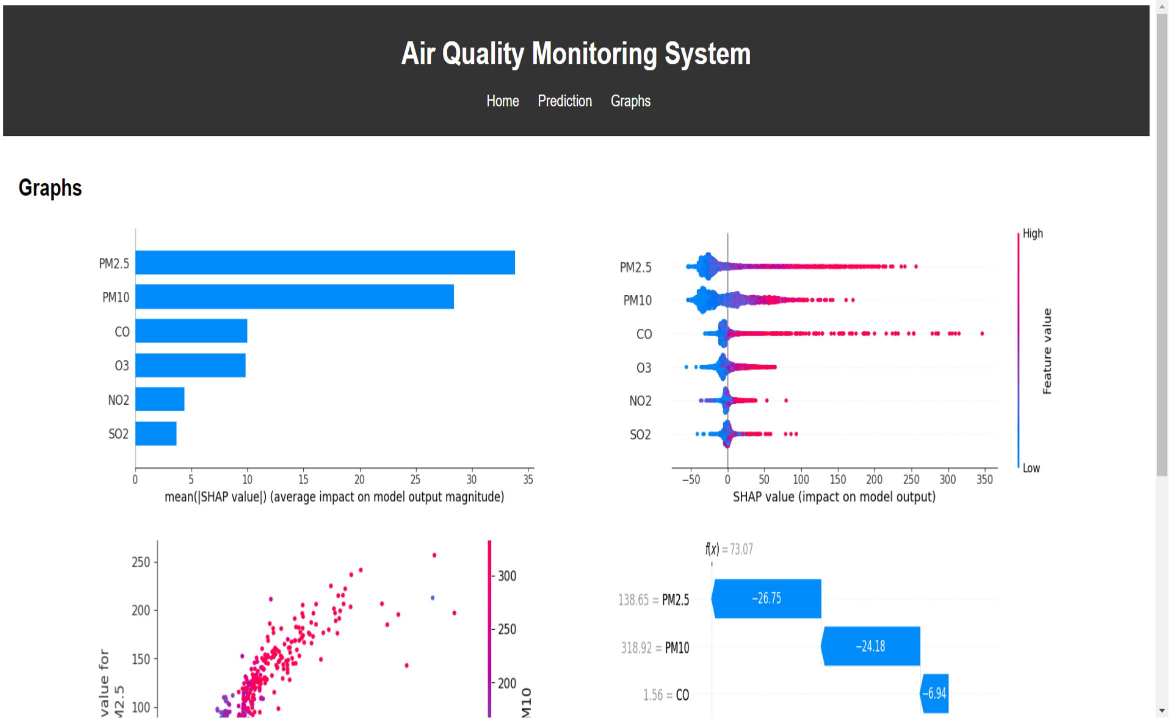



Figure 6.5: Website Graphs page-1

Figure 6.5 displays the graphs page of the website showing the various graphical representations of the influences of the features using explainable AI.

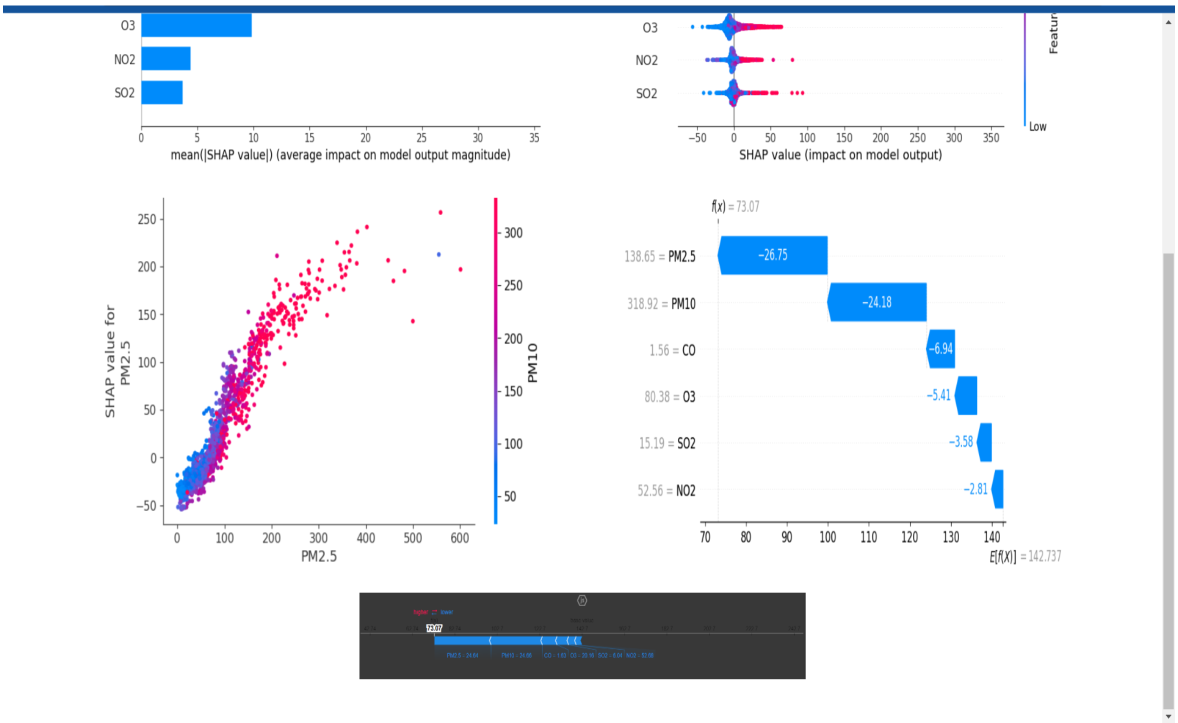



Figure 6.6: Website Graphs page-2

Figure 6.6 displays the graphs page of the website showing the summary plots, dependence plot, force plot and waterfall plot.

## 6.3 Quantitative Results

### 6.3.1 RMSE

Root Mean Squared Error (RMSE) is a valuable metric in regression analysis that provides a measure of the average magnitude of the errors between predicted and actual values.

### 6.3.2 MSE

Mean Squared Error (MSE) is a common evaluation metric used in regression analysis to measure the average of the squares of errors or deviations between predicted and actual values.This metric squares the errors before averaging, which penalizes larger errors more heavily than MAE.

### 6.3.3 R-squared

The coefficient of determination,is a widely used metric for evaluating the performance of regression models. It measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

```
LSTM - Mean Squared Error: 0.004398363098651445
LSTM - Root Mean Squared Error: 0.06632015605116928
LSTM - R-squared: 0.9012563123521581
```

Figure 6.7: Evaluation Metrics

Figure 6.7 is the RMSE, MSE and R-squared values of LSTM.The Mean Squared Error is found to be 0.00439. Root Mean Squared Error is 0.06632. The R-squared is 0.90125.The MSE and RMSE seem quite small, indicating that, on average, the model's predictions are close to the actual values.

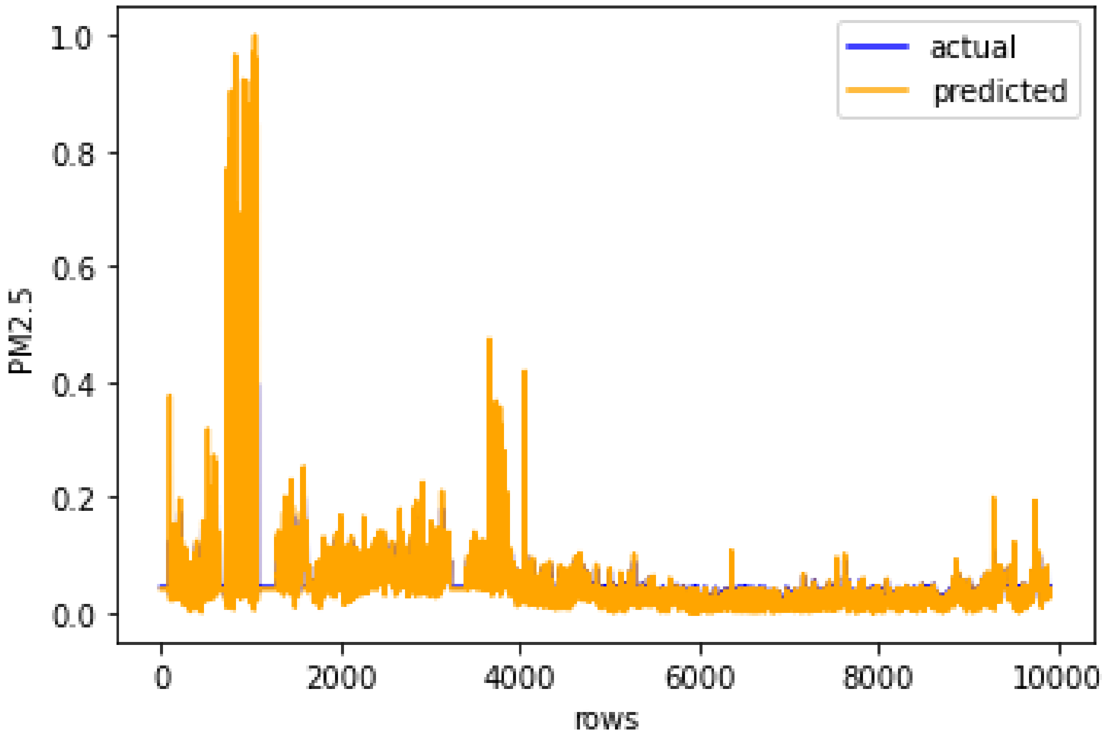



Figure 6.8: Actual vs Predicted

Figure 6.8 displays the actual vs predicted values of the dataset provided and we can see that it shows almost similar values.

### 6.3.4 IoT reading and Visualization

IoT is setup and readings from various sensors are stored in a cloud platform called ThingSpeak.ThingSpeak presents the data in graphical format, allowing users to analyze trends and patterns easily.

```
0. at command => AT+CIPMUX=1 OK
0. at command => AT+CIPSTART=0,"TCP","api.thingspeak.com",80 OK
0. at command => AT+CIPSEND=0,146 OK
0. at command => AT+CIPCLOSE=0 OK
CO2: 340.35 ppm
Methane: 406.51 ppm
NO2: 578.25 ppm
CO: 21.00 ppm
SO2: 21.00 ppm
Temperature: 35.00 °C
Humidity: 65.00 %
Ozone: 8.00 ppm
```

Figure 6.9: IoT readings

Figure 6.9 displays the readings and the connection requests messages shown in Arduino IDE

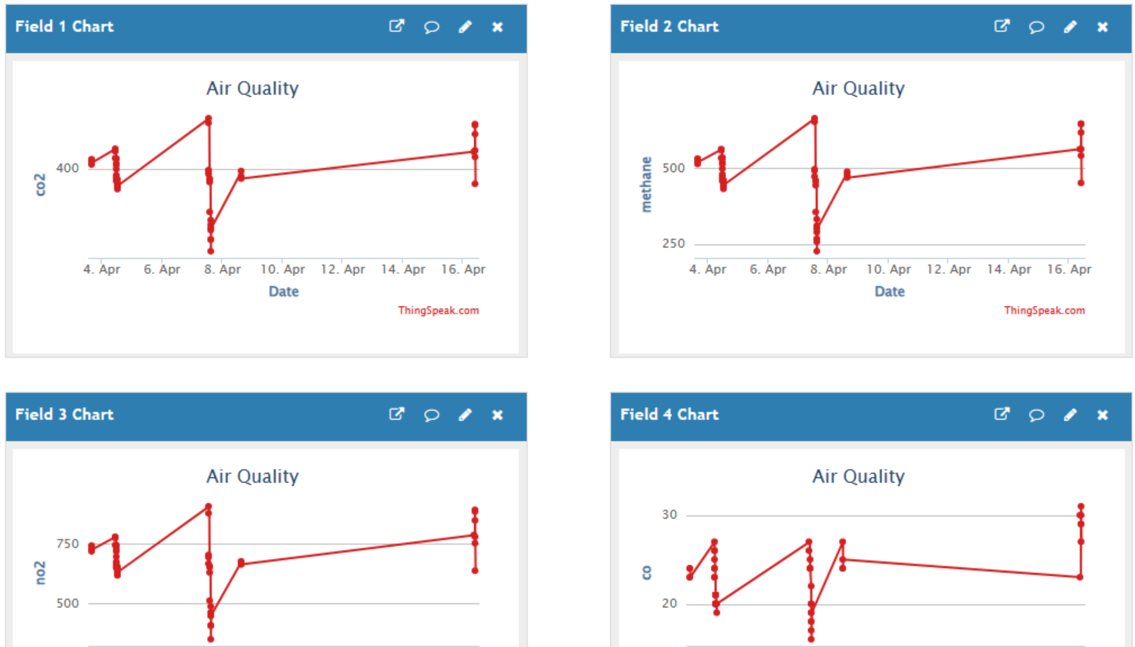



Figure 6.10: ThingSpeak interface

Figure 6.10 displays the real time IoT collected data of CO2,Methane,No2 and CO stored in ThingSpeak

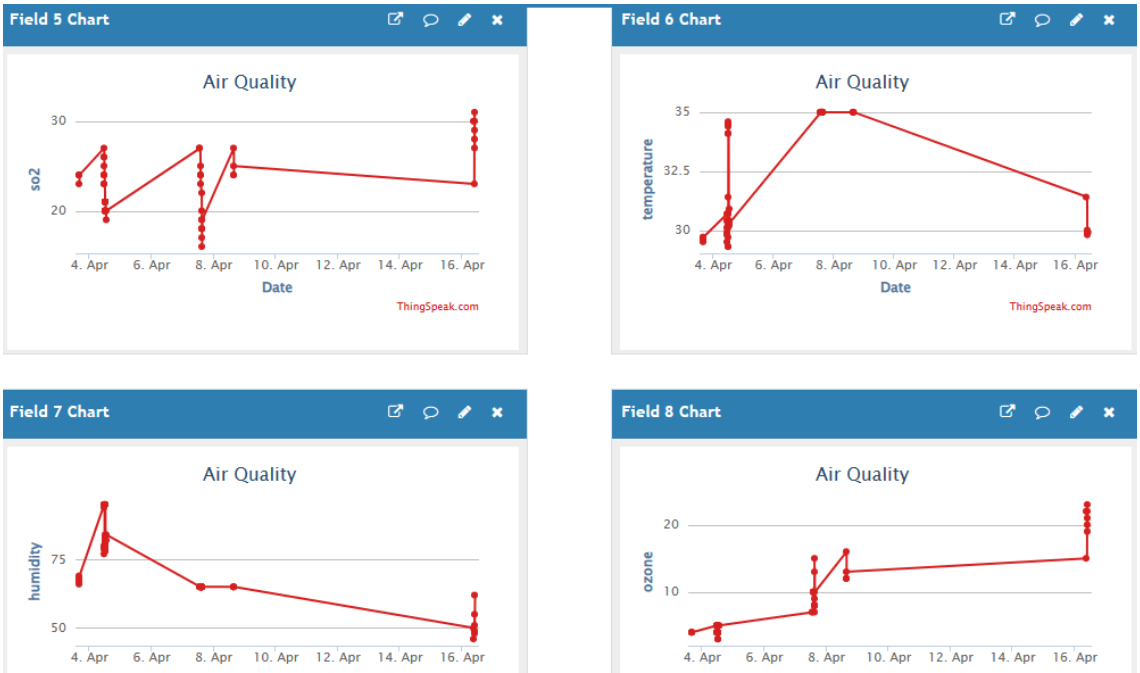



Figure 6.11: ThingSpeak interface

Figure 6.11 displays the real time IoT collected data of SO2,Temperature,Humidity and Ozone in ThingSpeak.

## 6.4 Conclusion

The air quality monitoring system is a powerful tool for keeping communities safe. By tracking different gases and comparing them to safe levels, we can spot potential health risks early. Our ability to predict future trends adds an extra layer of preparedness. This work highlights the importance of staying vigilant about our environment for everyone's well-being. We're committed to improving and expanding our monitoring efforts to ensure cleaner, healthier air for all.

# Chapter 7

# Conclusions & Future Scope

In summary, our project aimed to tackle the pressing issues of air quality prediction and disease forecasting through a comprehensive approach. By integrating historical datasets with real-time data from IoT sensors, we harnessed the power of Long Short-Term Memory (LSTM) models for accurate predictions. These models, designed to capture temporal dependencies in our environmental and health data, exhibited strong performance during training, validation, and testing. Additionally, our adoption of Shapely Additive Explanations (SHAP) provided transparency into the decision-making process of our air quality prediction model, shedding light on the individual contributions of features such as PM2.5, PM10, CO2, CH4, SO2, NO2, O3, temperature and humidity.

The insights gained from our project carry significant implications for environmental monitoring and public health management. Accurate air quality predictions empower authorities to implement timely interventions for pollution control, while disease warning system contribute to proactive healthcare strategies. The interpretability offered by SHAP enhances the trustworthiness of our models, fostering informed decision-making among stakeholders. As we look ahead, there is potential for further refinement by integrating additional data sources and advancing machine learning methodologies. Ethical considerations, including data privacy and model fairness, will remain central to the continued development and deployment of predictive models for environmental and health analytics.

The future scope of the project includes integrating additional data sources such as satellite imagery, exploring advanced machine learning models beyond LSTM, and developing real-time alert systems. Expanding the project's geographic scope, implementing continuous model monitoring, and fostering public engagement through accessible interfaces is also possible. It is possible to collaborate with healthcare institutions to enhance disease predictions and assess long-term environmental impacts.

# References


[1] Q. P. Ha, S. Metia and M. D. Phung, "Sensing Data Fusion for Enhanced Indoor Air Quality Monitoring," in *IEEE Sensors Journal*, vol. 20, no. 8, pp. 4430-4441, 15 April15, 2020, doi: 10.1109/JSEN.2020.2964396.

[2] X. Wang, J. Yan, X. Wang and Y. Wang, "Air Quality Forecasting Using the GRU Model Based on Multiple Sensors Nodes," in *IEEE Sensors Letters*, vol. 7, no. 7, pp. 1-4, July 2023, Art no. 6003804, doi: 10.1109/LSENS.2023.3290144.

[3] M. H. Nguyen, P. Le Nguyen, K. Nguyen, V. A. Le, T. -H. Nguyen and Y. Ji, "PM2.5 Prediction Using Genetic Algorithm-Based Feature Selection and EncoderDecoder Model," in IEEE Access, vol. 9, pp. 57338-57350, 2021, doi: 10.1109/ACCESS.2021.3072280.

[4] H. Chen, M. Guan and H. Li, "Air Quality Prediction Based on Integrated Dual LSTM Model," in IEEE Access, vol. 9, pp. 93285-93297, 2021, doi: 10.1109/ACCESS.2021.3093430.

[5] Dewen Seng, Qiyan Zhang, Xuefeng Zhang, Guangsen Chen, Xiyuan Chen, "Spatiotemporal prediction of air quality based on LSTM neural network", Alexandria Engineering Journal, Volume 60, Issue 2, 2021.

[6] Y. Yang, G. Mei and S. Izzo, "Revealing Influence of Meteorological Conditions on Air Quality Prediction Using Explainable Deep Learning," in IEEE Access, vol. 10, pp. 50755-50773, 2022, doi: 10.1109/ACCESS.2022.3173734.

[7] Rajasekar, Darshini and Sekar, Aravindh and Rajasekar, Magesh, Air Quality Monitoring and Disease Prediction Using IoT and Machine Learning (NOVEMBER 20, 2020). International Journal of Innovative Research in Computer Science and Technology (IJIRCST), Volume- 8, Issue- 6, November 2020.

# Appendix A: Presentation

# AirGuardian

Air Quality Analysis and Disease Warning System


Elizabath Gigi(U2003077)
Govind Prasad(U2003089)
Manu Jaison(U2003126)
Naina Fathima K(U2003140)

RSET


May 8, 2024

Guided by: Dr. Jisha G



## Problem Definition

The increasing air pollution in industrial and urban areas is getting worse due to transportation, fuel emissions, and other sources, so we need better systems to monitor and conduct air quality analysis.

## Project Objective

To enhance air quality prediction through the analysis of greenhouse gases, particulate matter and meteorological data, providing precise and timely forecasts to support healthier environments and informed decision-making.



## Novelty of Idea & Scope of Implementation

- The project involves monitoring greenhouse gas levels using IoT sensors to predict each gas concentration.
- The project employs explainable AI to understand the influence of all the features in Air quality prediction.
- The project provides risk assessments for potential diseases related to air pollution.

## Literature Review

| Reference Paper | Year | Insights |
|---|---|---|
| Air Quality Forecasting Using the GRU Model Based on Multiple Sensors Nodes | 2023 | **Model:**Single sensor,Multiple Sensor<br>**Advantages:**<br>• Lightweight GRU network<br>• Reduced dimensionality<br>**Disadvantages:**<br>• Complexity-performance trade-off<br>• Dependency on MI parameter setting |
| Revealing Influence of Meteorological Conditions on Air Quality Prediction Using Explainable Deep Learning | 2022 | **Model:**LSTM, GRU, SHAP<br>**Advantages:**<br>• effective environmental protection strategies can be formulated<br>• obtain the magnitude of the contribution of each type of feature to the prediction results<br>**Disadvantages:**<br>• high computational effort, large computational memory required, low computational efficiency |
| Air Quality Prediction using Graph Neural Network:A case study in Madrid | 2022 | **Model**:AT-GCN<br>**Advantages**:<br>• spatio-temporal dependencies are considered<br>• Multimodal data<br>**Disadvantages**:<br>• Only considers NO2 gas<br>• No real time data collection |



## Literature Review

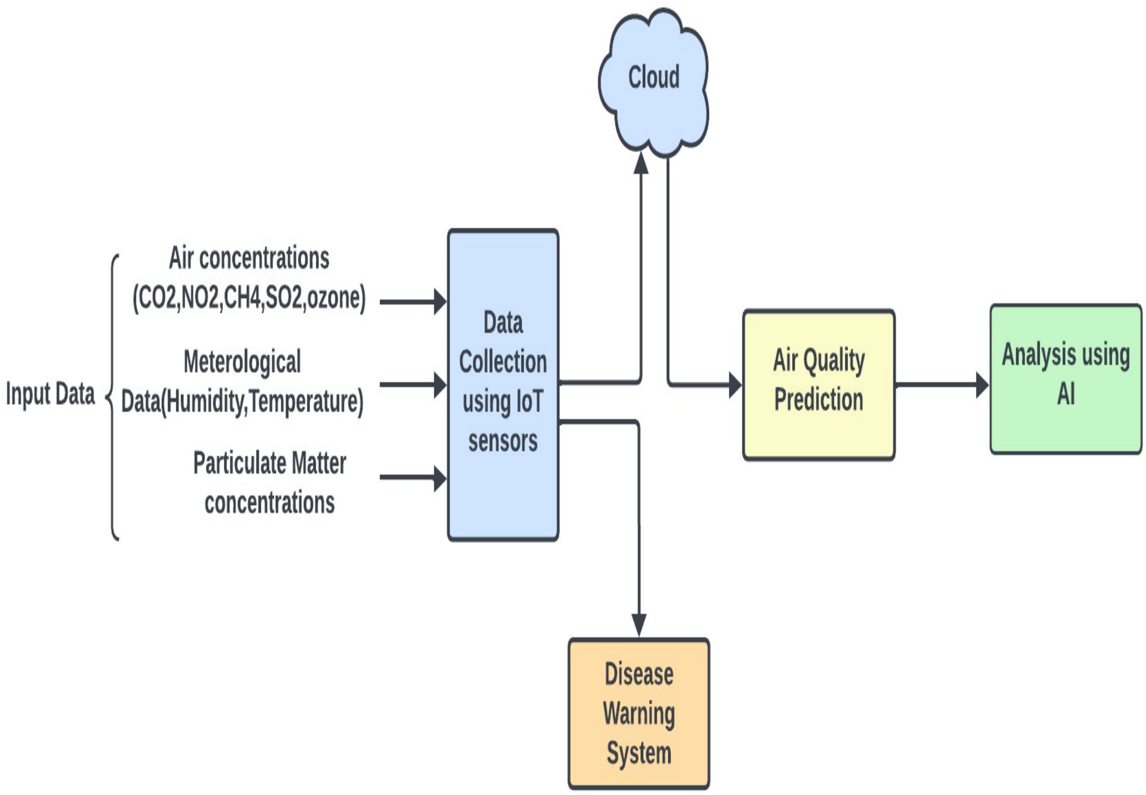
| | | |
|---|---|---|
| Air pollution prediction with machine learning: a case study of Indian cities | 2021 | **Model**:KNN,GNB,SVM, RF, XGBoost<br>**Advantages**:<br>• Uses SMOTE for solving data imbalance<br>• Compares 5 models<br>**Disadvantages**:<br>• Spatio-temporal dependencies are not considered<br>• Less accurate |
| Spatiotemporal prediction of air quality based on LSTM neural network | 2020 | **Model**:LSTM<br>**Advantages:**<br>• Spatio-temporal dependencies are considered<br>• Multi-Output and Multi-Index of Supervised Learning<br>**Disadvantages:**<br>• Concentrates on PM2.5<br>• No real time data collection |
| Air Quality Monitoring and Disease Prediction Using IoT and Machine Learning | 2020 | **Model**:XGBoost,Random Forest<br>**Advantages**:<br>• real time monitoring<br>• Data Accuracy<br>**Disadvantages**:<br>• Over dependence on technology<br>• Data overload |
| Real Time Localized Air Quality Monitoring and Prediction Through Mobile and Fixed IoT Sensing Network | 2020 | **Model:**Fixed Sensor,Hybrid<br>**Advantages:**<br>• Comprehensive spatial coverage<br>• Dynamic data collection<br>**Disadvantages:**<br>• Temporal sensitivity<br>• Limited applicability of machine learning |

## Methodology

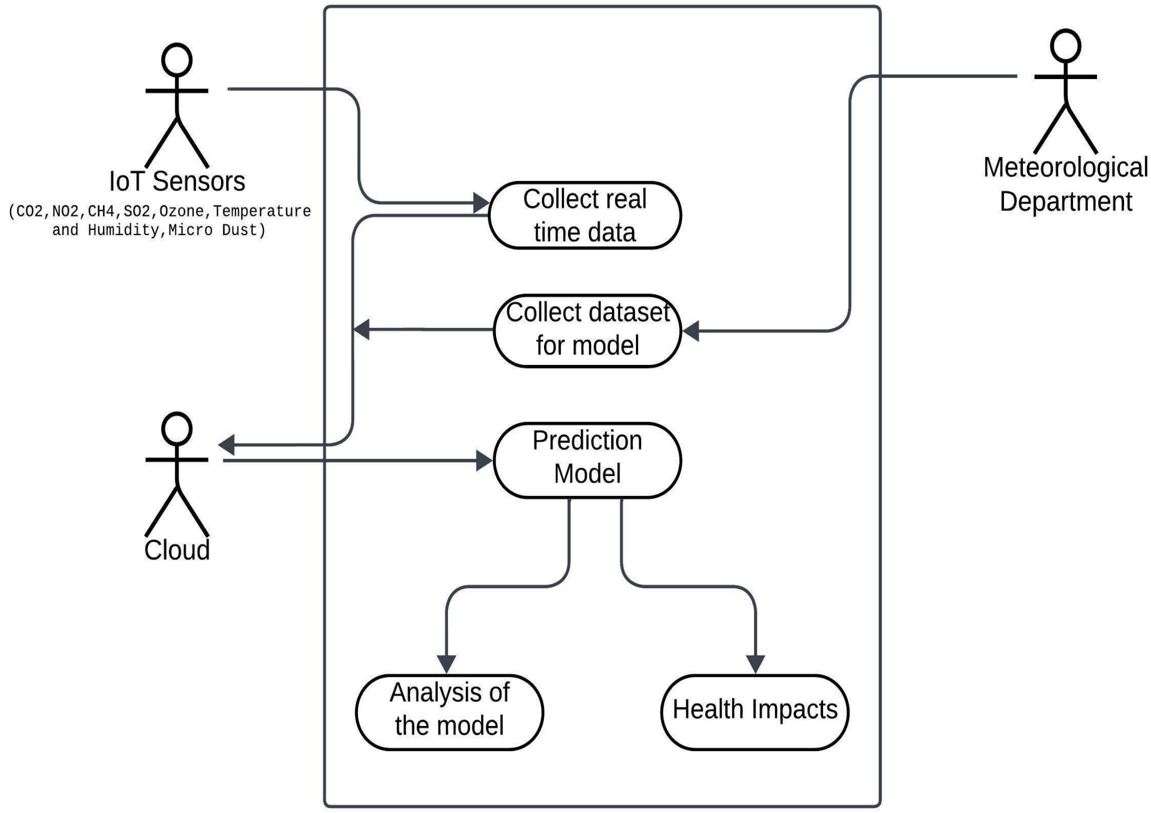
- This approach involves reading CO2, NO2,SO2,CH4,O3,temperature, humidity, and micro dust data using sensors. The dataset is then stored in a cloud server for comprehensive and accessible monitoring of air quality in real time.
- In Air quality Module, the dataset is preprocessed,normalized and trained to predict the air quality.
- The gas concentrations are measured and its impact on human health is predicted.
- We employ the explainable AI to reveal the impact of greenhouse gases and meteorological conditions on air quality prediction.



## Architecture Diagram



# Sequence Diagram





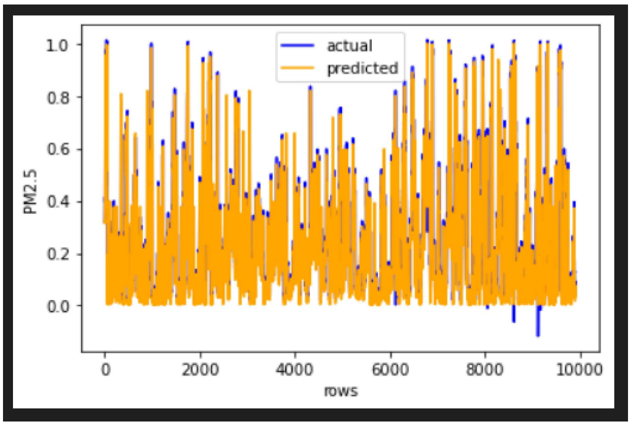

# Outputs

```
Message (Ctrl + Enter to send message to 'Arduino Uno' on 'COM3')

Temperature: 29.70 °C, Humidity: 72.00 %, MQ135 Value: 937, MQ7 Value: 939
Temperature: 29.70 °C, Humidity: 72.00 %, MQ135 Value: 936, MQ7 Value: 939
Temperature: 29.70 °C, Humidity: 71.00 %, MQ135 Value: 936, MQ7 Value: 939
Temperature: 29.70 °C, Humidity: 72.00 %, MQ135 Value: 936, MQ7 Value: 940
Temperature: 29.60 °C, Humidity: 72.00 %, MQ135 Value: 936, MQ7 Value: 939
Temperature: 29.60 °C, Humidity: 72.00 %, MQ135 Value: 936, MQ7 Value: 939
Temperature: 29.70 °C, Humidity: 72.00 %, MQ135 Value: 937, MQ7 Value: 940
Temperature: 29.60 °C, Humidity: 71.00 %, MQ135 Value: 936, MQ7 Value: 939
Temperature: 29.60 °C, Humidity: 72.00 %, MQ135 Value: 936, MQ7 Value: 939
Temperature: 29.60 °C, Humidity: 72.00 %, MQ135 Value: 936, MQ7 Value: 939
Temperature: 29.70 °C, Humidity: 72.00 %, MQ135 Value: 936, MQ7 Value: 939
Temperature: 29.70 °C, Humidity: 72.00 %, MQ135 Value: 937, MQ7 Value: 939
Temperature: 29.60 °C, Humidity: 72.00 %, MQ135 Value: 936, MQ7 Value: 939
Temperature: 29.60 °C, Humidity: 72.00 %, MQ135 Value: 936, MQ7 Value: 939
Temperature: 29.70 °C, Humidity: 72.00 %, MQ135 Value: 936, MQ7 Value: 939
Temperature: 29.70 °C, Humidity: 72.00 %, MQ135 Value: 936, MQ7 Value: 939
Temperature: 29.70 °C, Humidity: 72.00 %, MQ135 Value: 936, MQ7 Value: 939
```

IoT Sensor reading

# Outputs




LSTM

```
LSTM - Mean Squared Error: 0.004398363098651445
LSTM - Root Mean Squared Error: 0.06632015605116928
LSTM - R-squared: 0.9012563123521581
```

LSTM Evaluation Metrics



# Outputs

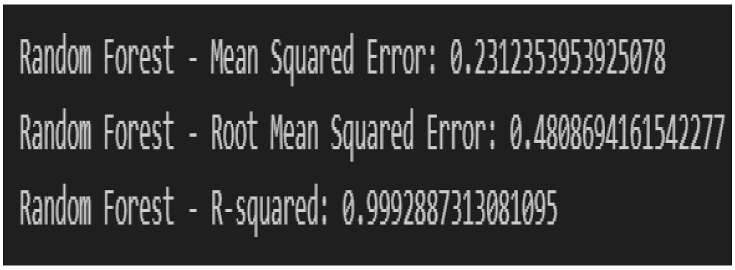
```
LinearRegression - Mean Squared Error: 8.185115213769079
LinearRegression - Root Mean Squared Error: 2.8609640357349964
LinearRegression - R-squared: 0.9748229885775535
```

Linear Regression

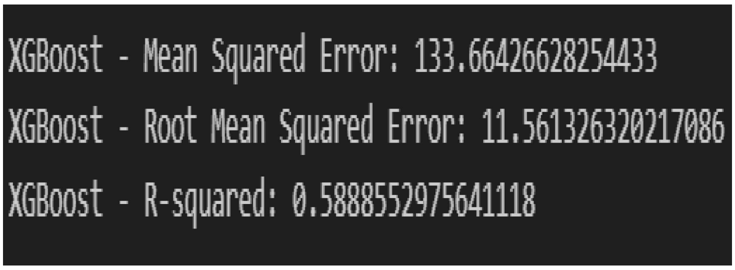
```
Decision Tree - Mean Squared Error: 0.5227852589589232
Decision Tree - Root Mean Squared Error: 0.7230389055638177
Decision Tree - R-squared: 0.9983919382815585
```

Decision Tree

# Outputs


```
Random Forest - Mean Squared Error: 0.2312353953925078
Random Forest - Root Mean Squared Error: 0.4808694161542277
Random Forest - R-squared: 0.9992887313081095
```

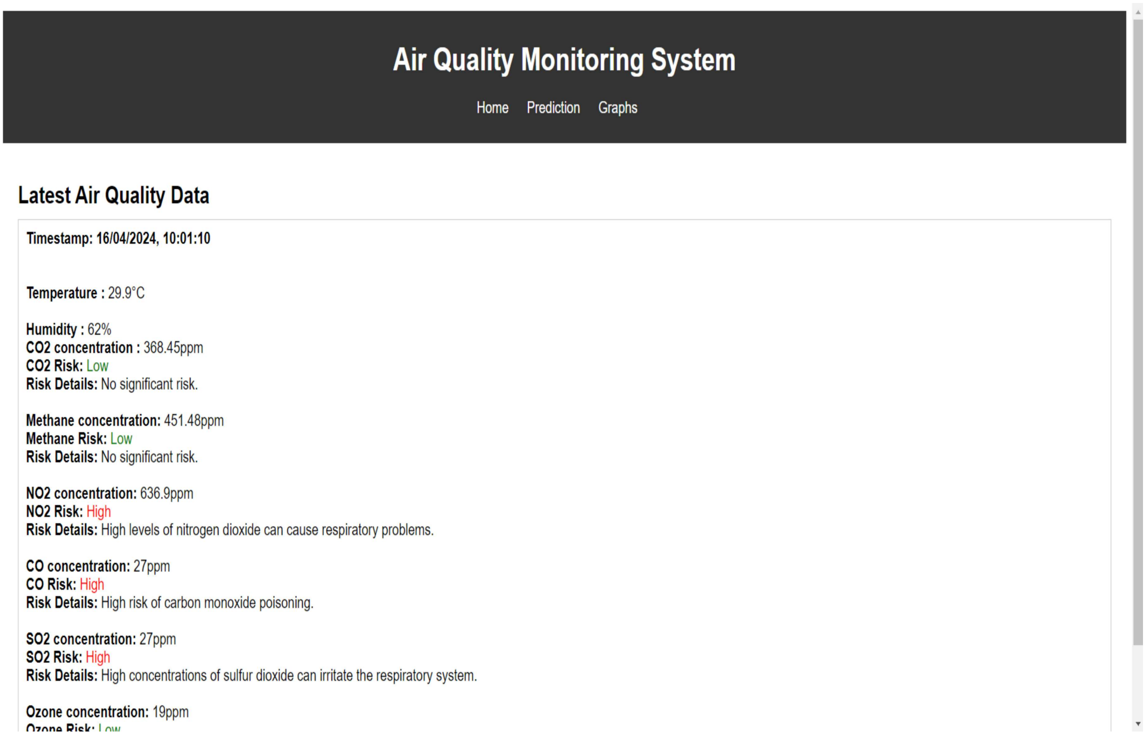
Random Forest


```
XGBoost - Mean Squared Error: 133.66426628254433
XGBoost - Root Mean Squared Error: 11.561326320217086
XGBoost - R-squared: 0.5888552975641118
```

XGBoost



# Outputs


Air Quality Monitoring System

Home Prediction Graphs

Latest Air Quality Data

**Timestamp: 16/04/2024, 10:01:10**

**Temperature :** 29.9°C

**Humidity :** 62%
**CO2 concentration :** 368.45ppm
**CO2 Risk:** Low
**Risk Details:** No significant risk.

**Methane concentration:** 451.48ppm
**Methane Risk:** Low
**Risk Details:** No significant risk.

**NO2 concentration:** 636.9ppm
**NO2 Risk:** High
**Risk Details:** High levels of nitrogen dioxide can cause respiratory problems.

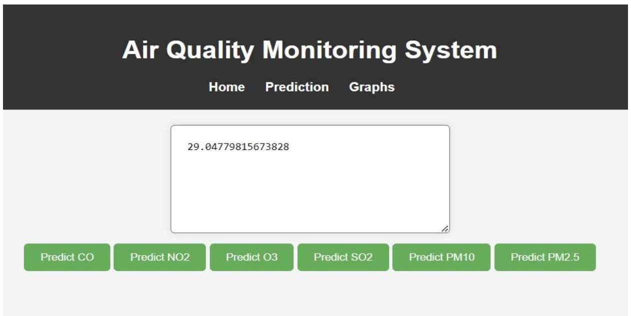
**CO concentration:** 27ppm
**CO Risk:** High
**Risk Details:** High risk of carbon monoxide poisoning.

**SO2 concentration:** 27ppm
**SO2 Risk:** High
**Risk Details:** High concentrations of sulfur dioxide can irritate the respiratory system.

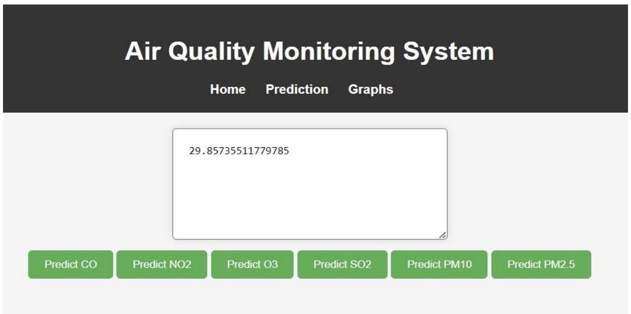
**Ozone concentration:** 19ppm
**Ozone Risk:** Low

Home Page

# Outputs



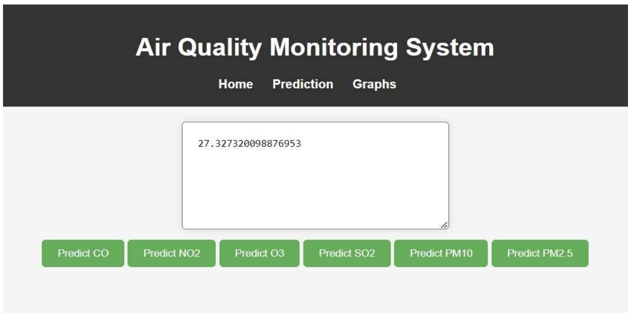

Prediction Page




Prediction Page

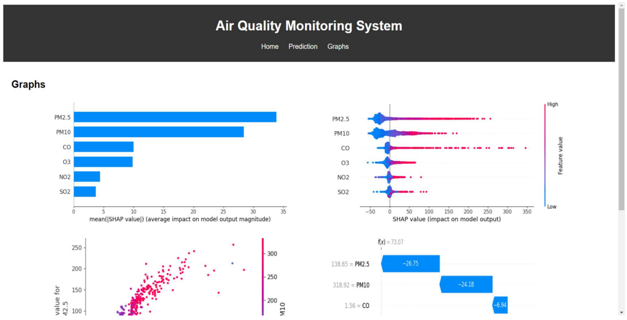


# Outputs

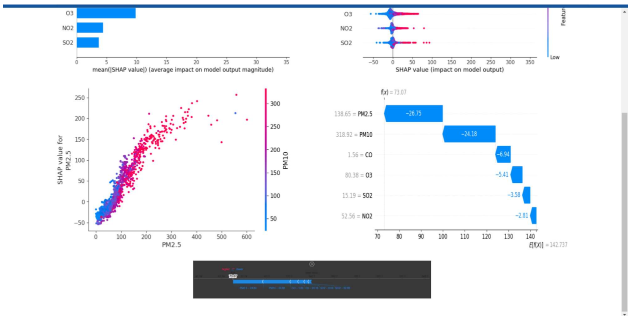



Prediction Page

## Outputs




Graph page



Graph Page



## Work Breakdown and Responsibilities

**Internet of Things: Naina Fathima**

**Air Quality Prediction: Elizabath Gigi**

**Disease Warning System & GUI: Manu Jaison**

**Analysis of Model: Govind Prasad**

## Conclusion

- In conclusion,we uses the data collected from the sensors to predict the air quality.
- The predicted air quality is analysed to understand the influence of various factors of air quality.



## Future Scope

- Incorporate additional sensors to capture more environmental parameters that could impact gas concentrations, such as humidity, temperature, and air quality indices.
- Implement real-time data streaming and processing techniques to enable instant alerts and notifications when gas concentrations exceed safe limits or when disease-related patterns are detected.
- Integrate geospatial data analysis to identify regional trends in gas concentrations and disease prevalence, providing insights for targeted interventions and policy-making.
- Develop a user-friendly mobile application that connects to the IoT system, allowing users to monitor gas concentrations, receive personalized health recommendations, and contribute data for community-wide analysis.

## Status of Paper Publication

The paper is based on our research on Air Quality Analysis and Health Warning systems.The research integrates IoT, Explainable AI, and predictive modeling to revolutionize air quality monitoring, forecasting air quality, explaining predictions, and estimating health risks, fostering public awareness and proactive health interventions.
The paper is communicated to the guide.



## Reference I


1. Y. Yang, G. Mei and S. Izzo, "Revealing Influence of Meteorological Conditions on Air Quality Prediction Using Explainable Deep Learning," in IEEE Access, vol. 10, pp. 50755-50773, 2022, doi: 10.1109/ACCESS.2022.3173734.
2. C. Liu, G. Pan, D. Song and H. Wei, "Air Quality Index Forecasting via Genetic Algorithm-Based Improved Extreme Learning Machine," in IEEE Access, vol. 11, pp. 67086-67097, 2023, doi: 10.1109/ACCESS.2023.3291146.
3. I. N. K. Wardana, S. A. Fahmy and J. W. Gardner, "TinyML Models for a Low-cost Air Quality Monitoring Device," in IEEE Sensors Letters, doi: 10.1109/LSENS.2023.3315249.

## Reference II


4. J. Han, H. Liu, H. Xiong and J. Yang, "Semi-Supervised Air Quality Forecasting via Self-Supervised Hierarchical Graph Neural Network," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 5, pp. 5230-5243, 1 May 2023, doi: 10.1109/TKDE.2022.3149815.
5. D. Iskandaryan, F. Ramos and S. Trilles, "Graph Neural Network for Air Quality Prediction: A Case Study in Madrid," in IEEE Access, vol. 11, pp. 2729-2742, 2023, doi: 10.1109/ACCESS.2023.3234214.
6. H. Kan, R. Chen, and S. Tong, "Ambient air pollution, climate change, and population health in China," Environ. Int., vol. 42, pp. 10–19, Jul. 2012.
7. H. Zhang, S. Wang, J. Hao, X. Wang, S. Wang, F. Chai, and M. Li, "Air pollution and control action in Beijing," J. Cleaner Prod., vol. 112, pp. 1519–1527, Jan. 2016.




## CO-PO & CO-PSO Mapping

**CO-PO AND CO-PSO MAPPING**

| | PO 1 | PO 2 | PO 3 | PO 4 | PO 5 | PO 6 | PO 7 | PO 8 | PO 9 | PO 10 | PO 11 | PO 12 | PSO 1 | PSO 2 | PSO 3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CS451.1 | | 1 | | 2 | | | | | | 2 | | | | | |
| CS451 .2 | | 2 | | | 2 | | | | | | | | 2 | | |
| CS451 | | 1.5 | | 2 | 2 | | | | | 2 | | | 2 | | |

**JUSTIFICATIONS FOR CO-PO MAPPING**

| Mapping | LOW/MEDIUM/HIGH | Justification |
|---|---|---|
| CS451.1 -P02 | L | Students do a literature survey while preparing for the seminar and project |
| CS451.1 -P04 | M | They reach valid conclusions after the literature survey |
| CS451.1- P010 | M | Seminar presentations help them to develop public speaking skills |
| CS451.2-P04 | H | They do detailed research in their area of interest which help them to analyze and synthesis data. |
| CS451.2-P05 | M | They understand the limitations of the existing techniques and can use the engineering techniques to arrive at valid conclusions |
| CS451.2-P010 | H | Writing seminar report help them to develop technical report writing skills. |
| CS451.2-PSO1 | M | By comparing different techniques, they can identify, analyze and design complex engineering problems. |

THANK YOU

# Appendix B: Vision, Mission, Programme Outcomes and Course Outcomes

# Vision, Mission, Programme Outcomes and Course Outcomes

**Institute Vision**

To evolve into a premier technological institution, moulding eminent professionals with creative minds, innovative ideas and sound practical skill, and to shape a future where technology works for the enrichment of mankind.

**Institute Mission**

To impart state-of-the-art knowledge to individuals in various technological disciplines and to inculcate in them a high degree of social consciousness and human values, thereby enabling them to face the challenges of life with courage and conviction.

**Department Vision**

To become a centre of excellence in Computer Science and Engineering, moulding professionals catering to the research and professional needs of national and international organizations.

**Department Mission**

To inspire and nurture students, with up-to-date knowledge in Computer Science and Engineering, ethics, team spirit, leadership abilities, innovation and creativity to come out with solutions meeting societal needs.

**Programme Outcomes (PO)**

Engineering Graduates will be able to:

**1. Engineering Knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

**2. Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

**3. Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

**4. Conduct investigations of complex problems**: Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

**5. Modern Tool Usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

**6. The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

**7. Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

**8. Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

**9. Individual and Team work**: Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.

**10. Communication**: Communicate effectively with the engineering community and with society at large. Be able to comprehend and write effective reports documentation. Make effective presentations, and give and receive clear instructions.

**11. Project management and finance**: Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work, as a member and leader in a team. Manage projects in multidisciplinary environments.

**12. Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change.

**Programme Specific Outcomes (PSO)**

A graduate of the Computer Science and Engineering Program will demonstrate:

**PSO1: Computer Science Specific Skills**
The ability to identify, analyze and design solutions for complex engineering problems in multidisciplinary areas by understanding the core principles and concepts of computer science and thereby engage in national grand challenges.

**PSO2: Programming and Software Development Skills**
The ability to acquire programming efficiency by designing algorithms and applying standard practices in software project development to deliver quality software products meeting the demands of the industry.

**PSO3: Professional Skills**
The ability to apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs thereby evolving as an eminent researcher and entrepreneur.

**Course Outcomes (CO)**

**Course Outcome 1:** Model and solve real world problems by applying knowledge across domains (Cognitive knowledge level: Apply).

**Course Outcome 2:** Develop products, processes or technologies for sustainable and socially relevant applications (Cognitive knowledge level: Apply).

**Course Outcome 3:** Function effectively as an individual and as a leader in diverse teams and to comprehend and execute designated tasks (Cognitive knowledge level: Apply).

**Course Outcome 4:** Plan and execute tasks utilizing available resources within timelines, following ethical and professional norms (Cognitive knowledge level: Apply).

**Course Outcome 5:** Identify technology/research gaps and propose innovative/creative solutions (Cognitive knowledge level: Analyze).

**Course Outcome 6:** Organize and communicate technical and scientific findings effectively in written and oral forms (Cognitive knowledge level: Apply).

# Appendix C: CO-PO-PSO Mapping

## CO-PO AND CO-PSO MAPPING

| | PO1 | PO2 | PO3 | PO4 | PO5 | PO6 | PO7 | PO8 | PO9 | PO10 | PO11 | PO12 | PSO1 | PSO2 | PSO3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CO1 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 3 | | |
| CO2 | 2 | 2 | 2 | | 1 | 3 | 3 | 1 | 1 | | 1 | 1 | | 2 | |
| CO3 | | | | | | | | | 3 | 2 | 2 | 1 | | | 3 |
| CO4 | | | | | 2 | | | 3 | 2 | 2 | 3 | 2 | | | 3 |
| CO5 | 2 | 3 | 3 | 1 | 2 | | | | | | | 1 | 3 | | |
| CO6 | | | | | 2 | | | 2 | 2 | 3 | 1 | 1 | | | 3 |

3/2/1: high/medium/low

## JUSTIFICATIONS FOR CO-PO MAPPING & CO-PSO MAPPING

| MAPPING | LOW/MEDIUM/ HIGH | JUSTIFICATION |
|---|---|---|
| 100003/ CS722U.1-PO1 | M | Knowledge in the area of technology for project development using various tools results in better modeling. |
| 100003/ CS722U.1-PO2 | M | Knowledge acquired in the selected area of project development can be used to identify, formulate, review |

| | | |
|---|---|---|
| | | research literature, and analyze complex engineering problems reaching substantiated conclusions. |
| 100003/ CS722U.1-PO3 | M | Can use the acquired knowledge in designing solutions to complex problems. |
| 100003/ CS722U.1-PO4 | M | Can use the acquired knowledge in designing solutions to complex problems. |
| 100003/ CS722U.1-PO5 | H | Students are able to interpret, improve and redefine technical aspects for design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions. |
| 100003/ CS722U.1-PO6 | M | Students are able to interpret, improve and redefine technical aspects by applying contextual knowledge to assess societal, health and consequential responsibilities relevant to professional engineering practices. |
| 100003/ CS722U.1-PO7 | M | Project development based on societal and environmental context solution identification is the need for sustainable development. |
| 100003/ CS722U.1-PO8 | L | Project development should be based on professional ethics and responsibilities. |
| 100003/ CS722U.1-PO9 | L | Project development using a systematic approach based on well defined principles will result in teamwork. |

| | | |
|---|---|---|
| 100003/ CS722U.1-PO10 | M | Project brings technological changes in society. |
| 100003/ CS722U.1-PO11 | H | Acquiring knowledge for project development gathers skills in design, analysis, development and implementation of algorithms. |
| 100003/ CS722U.1-PO12 | H | Knowledge for project development contributes engineering skills in computing & information gatherings. |
| 100003/ CS722U.2-PO1 | H | Knowledge acquired for project development will also include systematic planning, developing, testing and implementation in computer science solutions in various domains. |
| 100003/ CS722U.2-PO2 | H | Project design and development using a systematic approach brings knowledge in mathematics and engineering fundamentals. |
| 100003/ CS722U.2-PO3 | H | Identifying, formulating and analyzing the project results in a systematic approach. |
| 100003/ CS722U.2-PO5 | H | Systematic approach is the tip for solving complex problems in various domains. |
| 100003/ CS722U.2-PO6 | H | Systematic approach in the technical and design aspects provide valid conclusions. |

| | | |
|---|---|---|
| 100003/ CS722U.2-PO7 | H | Systematic approach in the technical and design aspects demonstrate the knowledge of sustainable development. |
| 100003/ CS722U.2-PO8 | M | Identification and justification of technical aspects of project development demonstrates the need for sustainable development. |
| 100003/ CS722U.2-PO9 | H | Apply professional ethics and responsibilities in engineering practice of development. |
| 100003/ CS722U.2-PO11 | H | Systematic approach also includes effective reporting and documentation which gives clear instructions. |
| 100003/ CS722U.2-PO12 | M | Project development using a systematic approach based on well defined principles will result in better teamwork. |
| 100003/ CS722U.3-PO9 | H | Project development as a team brings the ability to engage in independent and lifelong learning. |
| 100003/ CS722U.3-PO10 | H | Identification, formulation and justification in technical aspects will be based on acquiring skills in design and development of algorithms. |
| 100003/ CS722U.3-PO11 | H | Identification, formulation and justification in technical aspects provides the betterment of life in various domains. |
| 100003/ CS722U.3-PO12 | H | Students are able to interpret, improve and redefine technical aspects with mathematics, science and |

| | | |
|---|---|---|
| | | engineering fundamentals for the solutions of complex problems. |
| 100003/ CS722U.4-PO5 | H | Students are able to interpret, improve and redefine technical aspects with identification formulation and analysis of complex problems. |
| 100003/ CS722U.4-PO8 | H | Students are able to interpret, improve and redefine technical aspects to meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations. |
| 100003/ CS722U.4-PO9 | H | Students are able to interpret, improve and redefine technical aspects for design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions. |
| 100003/ CS722U.4-PO10 | H | Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools for better products. |
| 100003/ CS722U.4-PO11 | M | Students are able to interpret, improve and redefine technical aspects by applying contextual knowledge to assess societal, health and consequential responsibilities relevant to professional engineering practices. |
| 100003/ CS722U.4-PO12 | H | Students are able to interpret, improve and redefine technical aspects for demonstrating the knowledge of, and need for sustainable development. |

| | | |
|---|---|---|
| 100003/ CS722U.5-PO1 | H | Students are able to interpret, improve and redefine technical aspects, apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice. |
| 100003/ CS722U.5-PO2 | M | Students are able to interpret, improve and redefine technical aspects, communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions. |
| 100003/ CS722U.5-PO3 | H | Students are able to interpret, improve and redefine technical aspects to demonstrate knowledge and understanding of the engineering and management principle in multidisciplinary environments. |
| 100003/ CS722U.5-PO4 | H | Students are able to interpret, improve and redefine technical aspects, recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change. |
| 100003/ CS722U.5-PO5 | M | Students are able to interpret, improve and redefine technical aspects in acquiring skills to design, analyze and develop algorithms and implement those using high-level programming languages. |
| 100003/ CS722U.5-PO12 | M | Students are able to interpret, improve and redefine technical aspects and contribute their engineering skills in |

| | | |
|---|---|---|
| | | computing and information engineering domains like network design and administration, database design and knowledge engineering. |
| 100003/ CS722U.6-PO5 | M | Students are able to interpret, improve and redefine technical aspects and develop strong skills in systematic planning, developing, testing, implementing and providing IT solutions for different domains which helps in the betterment of life. |
| 100003/ CS722U.6-PO8 | H | Students will be able to associate with a team as an effective team player for the development of technical projects by applying the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems. |
| 100003/ CS722U.6-PO9 | H | Students will be able to associate with a team as an effective team player to Identify, formulate, review research literature, and analyze complex engineering problems |
| 100003/ CS722U.6-PO10 | M | Students will be able to associate with a team as an effective team player for designing solutions to complex engineering problems and design system components. |
| 100003/ CS722U.6-PO11 | M | Students will be able to associate with a team as an effective team player, use research-based knowledge and research methods including design of experiments, analysis and interpretation of data. |

| | | |
|---|---|---|
| 100003/ CS722U.6-PO12 | H | Students will be able to associate with a team as an effective team player, applying ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice. |
| 100003/ CS722U.1-PSO1 | H | Students are able to develop Computer Science Specific Skills by modeling and solving problems. |
| 100003/ CS722U.2-PSO2 | M | Developing products, processes or technologies for sustainable and socially relevant applications can promote Programming and Software Development Skills. |
| 100003/ CS722U.3-PSO3 | H | Working in a team can result in the effective development of Professional Skills. |
| 100003/ CS722U.4-PSO3 | H | Planning and scheduling can result in the effective development of Professional Skills. |
| 100003/ CS722U.5-PSO1 | H | Students are able to develop Computer Science Specific Skills by creating innovative solutions to problems. |
| 100003/ CS722U.6-PSO3 | H | Organizing and communicating technical and scientific findings can help in the effective development of Professional Skills. |