Air Pollution Prediction System in Smart Cities Using Data Mining Technique

Debopriya Manna[1], Rohan Mondal[1], Arpan Sanyal[1], Ahana Biswas[1], Hritam Roy[1] and Subhajyoti Barman[2]

[1]B. Tech, Dept. of Computer Science and Engineering, Techno India University, Kolkata, West Bengal, India.
[2] Assistant Professor, Dept. of Computer Science and Business Systems, Techno India University, Kolkata, West Bengal, India.

---***---

Abstract - *In every nation, Air Pollution has grown to be a serious problem. Among mankind's most serious issues, pollution in the air is one of the major contributors to health and climate change. Fears over rising air pollution, that could harm human health, the health of every living creature, and the advancement of the world economic growth, are shared by both the government and the general population. For this reason, air pollution forecasting has become crucial. To address this issue, several deep learning models were applied. AQI was calculated to conduct this study. Various pollution-causing gases like nitrogen dioxide, sulphur dioxide, carbon monoxide and ozone and particulate matter like PM10 and PM2.5 were studied. The vast quantity and diversity of data collected by air pollution monitoring stations across different cities have made air pollution forecasting an important topic. This research incorporates an LSTM (Long-Short Term Memory), ARIMA, Prophet Linear Regression, and Polynomial Regression. The dataset is primarily comprised of different pollution-causing components data collected from the Central Pollution and Control Board (CPCB). Our project aims to investigate the weather in several Indian cities and record their AQI level. Using additional variables, we attempt to determine the extent of air pollution. Eventually, we base our findings on the historical pattern of the graph. Based on the previous year's pollution data, we attempt to forecast how the weather will evolve over the following several years. Different time series models were studied and the best-suited model was decided based on different parameters. This paper utilises ARIMA, SARIMA, Prophet, Long short-term memory (LSTM), Linear Regression and Polynomial Regression. These models aim to find the future trend for the upcoming months. Root Mean Square Error (RMSE) was used as performance metrics to evaluate the models, along with it Mean Absolute Percentage Error (MAPE) was also utilised to assess the models.*

***Key Words*: LSTM, ARIMA, RMSE, MAPE, AQI.**

1. INTRODUCTION

In the last few years, there have been constant changes in the environment which led to degrading air quality due to the presence of various harmful air pollutants. Air pollution is the presence of harmful pollutants in the air that are contaminating the indoor and outdoor air. Health and climate are being compromised due to air pollution. It is becoming a major threat to the environment. The major contributor to air pollution is fine particulate matter which is leading to lung cancer, acute and chronic respiratory disorders, heart attacks, and strokes.

Vehicles, domestic energy use for heating and cooking, and other sources are the main contributors to outdoor pollution, power generation, agriculture/waste incineration, and industry. The primary air pollutants are PM 10 with a diameter of less than 10 microns and PM 2.5 is more hazardous as it has a diameter of less than 2.5 microns. Sulphur Dioxide (SO2) and PM 2.5 are caused due to unburned fuel and also by processed byproducts. Due to fuel combustion Nitrogen Dioxide (NO2), Ozone (O3) and Carbon Monoxide (CO) are produced.

CO is the most dangerous one and it is also known as a silent killer. It deprives the brain and heart of oxygen that is required by the body to function, by entering our blood cells directly and replacing the oxygen in our body. When there is an increase in the pollutant levels it leads to humans losing consciousness and vomiting. When the exposure to these harmful pollutants is too long, it can eventually damage the brain cells in the body or can cause death. The government of India uses PM 10 and PM 2.5 as the major criteria for Air Quality Index (AQI) calculation.

The calculation of AQI comprises a minimum of three parameters out of which one must be either PM 2.5 or PM 10. The calculation of sub-indices requires 16 hours of data. For the Calculation of AQI, the Sub-indices for individual pollutants at a monitoring location are calculated using their 24-hour average concentration value (8 hours in the case of CO and O3) and health breakpoint concentration range. The sub-index that will be worst, is the AQI for that location.

Equation :

$I_{si} = [((C_{obs} – C_{min}) (I_{max} – I_{min})) / C_{max} – C_{min}] + I_{min}$

Where,

I_{si} = Sub-index value of observed pollutant

C_{obs} = Observed pollutant concentration

C_{max} = Maximum concentration of AQI breakpoint that contains $\leq$ C_{obs}

C_{min} = Minimum concentration of AQI breakpoint that contains C_{obs}

I_{max} = Maximum AQI value corresponding to $\leq$ C_{obs}

I_{Lo} = Minimum AQI value corresponding to C_{min}

1.1 Motivation

A silent killer is taking away the lives of millions, yet most people remain unaware of its danger. World Health Organisation (WHO) has been warning us regarding this threat but it usually remains unnoticed among the masses. Approximately 7 million premature deaths are caused because of this issue. Among these deaths, 55% are caused by heart failure, 21% by respiratory infection and the other 24% by other lung diseases. In addition to these certain psychological disorders such as Alzheimer's are also affecting some groups of people. Poor air quality significantly increases mental health risks like depression by 50%, personality disorders by 162%, and schizophrenia by 148%. According to the IQAir website among the top 50 polluted cities globally, 42 of them are in India. Research from various cancer institutes shows that while most lung cancer patients a decade ago were smokers, now 50% of diagnosed patients have never smoked.

The WHO studied the indoor air quality of several severely polluted areas and discovered something in common which was the use of Mud and Kerosene stoves. The use of these stoves significantly increases the emission of CO2, CO, Sulphur Oxides and Nitrogen Oxides, leading to higher pollution levels indoors. This has become a significant issue at the small-scale level itself. Attention must be given to these small-scale sources because addressing them is crucial, thus solutions must start at home, by tackling indoor air pollution first and further can create a ripple effect that raises awareness and drives larger public initiatives. Only then we can build a comprehensive approach to improve air quality and safeguard public health.

1.2 Research Question

This thesis will be concerned with the following research questions:

RQ 1. "How can Data Mining Techniques develop air pollution prediction systems in smart cities?"

RQ 2. "How can it be beneficial in providing better health to humans by forecasting AQI?"

RQ 3. "How can an IOT-based device which is fitted with suitable sensors help in bringing down air pollution on a small-scale level?"

RQ 4. "How will the entire model help the general public in the future?"

1.3 Objectives

Objectives for this study are as follows

- To analyse studies related to AQI.
- To decide the most suitable time series algorithm used to predict the AQI.
- To evaluate the performance and accuracy of different models and discover the most suitable time series algorithm for predicting AQI.
- To build a model for a small-scale radius to combat indoor air pollution, detect AQI and provide preventive measures.

2. Literature Review

This section comprises information about research papers related to the Air Pollution Prediction System.

This paper comprises air quality prediction which uses deep learning. It emphasizes the Long Short-Term Memory (LSTM) model. It had a goal to analyse the weather conditions in various cities across India and to correlate them with observed AQI values. It helps to understand the extent of air pollution in different regions by searching various factors that contribute to it. It uses historical weather data, which is further attempted to predict future weather patterns and possible effects on the quality of air. In the beginning, statistical models such as VARMA were used to analyse the trends, but after encountering limitations, the project was further switched to the LSTM model to achieve better performance. Once the improvements were seen in the predictive capabilities when machine learning algorithms were used in place of previous approaches. Applying the methods, helped to provide insights into the AQI levels of various cities across India. Depending on the error values obtained from their predictive models for the cities of Kolkata, Bangalore, and Hyderabad it was found that the LSTM model outperforms the other methods when it comes to Mean Squared Error (MSE) for predicting AQI values.

- In the case of Kolkata, the MSE for the VAR model is 81.52, for VARMA it is 76.911, and for LSTM it is 22.94.
- For Bangalore, the MSE for VAR is 82.03, for VARMA is 73.23, and for LSTM it is 14.94.

- In Hyderabad, the MSE for VAR is 85.63, for VARMA it is 51.61, and for LSTM it is 13.12.

The decreased MSE values indicate better predictive performance, and for the above case, LSTM consistently demonstrated the lowest MSE across all three cities. Thus, it suggests that the LSTM model is more effective in predicting AQI values accurately and also captures the correlations within the data. [1]

The following study consists of Recurrent Neural Network (RNNs) and Long Short-Term Memory (LSTM) units. It was used to draw insights from the time series data on air quality and also meteorological conditions. This paper also aims to understand the advanced research on urban air quality and thus helps governments in formulating beneficial policies by providing valuable information. Further, the forecast of PM2.5 and AQI values presents several challenges as the influence of various factors on air pollution thus making it difficult to discern past repetitive patterns. At first, the model was trained only on the first part of the dataset comprised of values from January 2015 to June 2017 and then the learned weights were transferred to the second part which comprised data from June 2017 to March 2018. The testing part was conducted on the data collected from February 2018 to March 10th,2018. In the case of the Seoul dataset, it contained complete information from January 2008 to April 2018 and it consisted of over 2 million hourly records across 25 districts. This model was trained on data from 2008 to 2016, and the testing part was performed on data from January 2017 to April 2018. Twenty per cent of the training set was randomly set aside as the validation set. The joint of multiple RNN layers, Mean Absolute Error (MAE) loss function, and transfer method yield outstanding prediction results, with MAE proving more effective than Mean Squared Error (MSE). The depicted model demonstrates significant results that predict PM2.5 and AQI based on historical meteorological data thus contributing to policy-making and resource allocation efforts., Mean Absolute Error (MAE) loss function and transfer method yield outstanding prediction results. Thus, proving that MAE yields more effective results than MSE. The Test Root Mean Squared Error (RMSE) for 2017-2018 with different settings shows the performance scores for different model settings over time intervals ranging from 8 hours to 24 hours. Two datasets are utilized: Joint Dataset and Seoul Dataset, alongside variations in model architectures: RNNs and RNN. In the case of the Combined Dataset, models incorporating RNNs show an increase in scores from 26.27 at 8 hours to 31.29 at 24 hours. In contrast, the RNN variant starts slightly lower at 25.93 and follows a similar upward trend, reaching 31.52 after 24 hours. In contrast, using the Seoul Dataset with RNNs produces higher initial scores than the Joint Dataset, starting at 27.51 and increasing to 31.8 at 24 hours, indicating a consistent improvement over time. Overall, the presented model provides promising results for predicting PM2.5 and AQI using historical meteorological data, with potential applications in policymaking and resource allocation. In the future, more evaluation and refinement of the model are required to predict the accuracy for future applications. [2]

This paper focuses on using Time Series Analysis of air quality indicators which is used to forecast air pollutant concentrations. This case specifically focuses on SO2, NO2, and PM10 and focuses on the cities of Margao and Sanguem located in Goa, India. The research was conducted using time series forecasting which used Auto-Regressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) models. The performance of these models was measured. This paper used Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE). When the comparison between both ARIMA and LSTM were done results were found that prediction using the LSTM model gives the air pollution concentration more accurately. It was also found that the MAPE and RMSE performance measures show that the LSTM method outperforms the ARIMA model. By taking an example from the above paper it is found in Margao, that the ARIMA model yields the highest MAPE of 15% for PM10 air pollutants. In contrast, for NO2 air pollutants in Sanguem, both ARIMA and LSTM models produce a MAPE of 3.3%. The following are the MAPE and RMSE values for each algorithm across various cities and pollutants:

- Margao (SO2): Using ARIMA MAPE = 5.7304 and RMSE = 0.6444; Using LSTM MAPE = 3.6485 and RMSE = 0.3697.

- Margao (NO2): Using ARIMA MAPE = 7.3615 and RMSE = 0.9967; Using LSTM MAPE = 6.1957 and RMSE = 0.8001

- Margao (PM10): Using ARIMA MAPE = 15.6621 and RMSE = 6.577; Using LSTM MAPE = 9.3682 and RMSE = 4.5959

- Sanguem (SO2): Using LSTM MAPE = 9.6743 and RMSE = 0.9534; Using LSTM MAPE = 6.5892 and RMSE = 0.6964

- Sanguem (NO2): Using LSTM MAPE = 3.3521 and RMSE = 0.5688; Using LSTM MAPE = 3.3661 and RMSE = 0.5784

- Sanguem (PM10): Using LSTM MAPE = 5.7695 and RMSE = 3.3945; Using LSTM MAPE = 3.9679 and RMSE = 3.1323

Overall, the above values depict the accuracy of each algorithm in predicting the pollutant concentrations, with lower MAPE and RMSE values which suggest better performance. In most cases, the LSTM model outperforms the ARIMA model which demonstrates its effectiveness in forecasting air pollutant concentrations. [3]

The following research comprises of Long Short-Term Memory model which is a powerful type of Recurrent Neural Network (RNN). It is used to predict sequential data in several fields like stock, market trends, etc. About air quality prediction, LSTM models can forecast AQI levels using historical data from weather station sensors. The LSTM's operational mechanism consists of three key gates. They are the input gate, the forget gate, and the output gate. These gates help to control the flow of information within the LSTM unit. It further decides which data to keep, discard, or send to other units in the network. The proposed system trains the LSTM model with the use of historical air quality data from a weather station in Beijing, China. The dataset in this project is divided into training and testing sets. After the division preprocessing techniques are used before the data is fed into the above-mentioned model. The Network parameters are set that include the optimizer, which is critical for improving neural network performance. The Adam optimizer which helps in adaptability and efficiency in determining learning rates, is used in this study. Multiple pollutants contribute to air pollution and influence PM 2.5 concentration. The optimizer was set to provide different learning rates for each pollutant feature, thereby improving the model's predictive accuracy. The Mean Square Error (RMSE) metric is used to calculate errors, and the proposed model had an error of around 0.1. The paper introduces a robust approach for estimating pollutant concentration levels quickly and accurately. It ensures a wide range of applications, which includes weather stations and user-centric platforms. Notably, the projected PM10 concentration for the next hour is 27.149012, with a mean square error of 0.0009564670581444191. The model showed impressive performance metrics, with a training score of 0.00004 and a testing score of 0.00017.[4]

Air pollution is one of the most important environmental issues regarding the health of public and the nature worldwide. It results from various reasons like industrial activities, vehicle emissions and smoke released by fossil fuel burning. Air pollution results in climate change and respiratory health problems, cardiovascular diseases and even can lead to cancer. The IOT model proposed here uses Raspberry Pi which relates to the Arduino Uno microcontroller. Since Raspberry Pi 2 model B does not have inbuilt WIFI, so WIFI adapter is used. DSM501A is used as a Particulate matter sensor. DHT22 and BMP180 temperature, pressure and humidity sensors are used. MQ9 and MQ135 sensors are used for gas concentration detection. The data is sent over the MQTT server and accessed by the client. The system exhibits medium to low cost, low power, compact and high accuracy.[5]

As proposed here, Smart Air is to collect accurate indoor air quality data. STM 32 F407IG is used here as the microcontroller. Laser Dust sensor PM2007 is used. This sensor can detect specifically PM 2.5 and PM 10. GSBT11-P110 volatile organic compound sensor is used to detect hydrocarbon-based petroleum products in the air. RCU890L LTE acts as the networking module. Here the data is stored using MySQL database. Based on the results air quality was accessed and people around were notified to improve the air quality. The proposed system achieved many merits like, accurate monitoring and alert system. It enhanced the security of data and platforms using Amazon web services. It also shows an expandable interface, where installation and addition of new sensors is easy. [6]

The proposed model has MQ2, MQ4 and MQ135 sensors of the MQ series gas sensors where Tin-Oxide (SnO_2) is used as a sensing element. GP2Y1010AU0F an optical dust sensor was used. It has a sensitivity of up to 0.5 mg/m^3. The software setup comprises three parts. Firstly, a C program is uploaded to the memory of the Arduino controller for the collection and transmission of data. Then an Android application receives the data and stores it in a CSV file and a cloud service receives it. In this study, the hourly data has been collected at two different stations in the Delhi-NCR region upon which time series machine learning models are used to predict and forecast the future AQI. [7]

In this model, the esp8266 module is used as the IOT gateway. The System proposes to measure indoor temperature, pressure, and air quality and make the user aware of any undesirable change in the environment leading to problems in their lives. The collected data will be analysed and data patterns can be extracted. Here, DHT11 and K-30 infrared CO_2 sensors are used. The esp8266 module is low in cost, reliable, uses less power and can send data over TCP/IP network. The model can send messages and alert people too. [8]

The survey deals with the capacity to analyse and make choices based on the current activity circumstance in the city given by the progressing checking of contaminants. Past enquiry has utilized a few AI gadgets to foresee contamination; in any case, comparative examination of these procedures is regularly fundamental to get a way better understanding of their preparing times for different datasets. Pre-processing the time arrangement is a portion of this method. They have explored the utilization of Energetic Time Distorting, LSTM, and ARIMA for time series forecast; in any case, pre-processing involves a similitude degree. Kmeans and Back Vector Relapse are at that point utilized to categorize the spatio-temporal contamination information of different places over ten years. The results illustrated that Kmeans clustering taken after by ARIMA can be utilized for time arrangement desires. Energetic Time Distorting is a way better measure to calculate the arrangement between two time series. In the long final, ARIMA can be utilized to make time-series relapse over bunches for a figure. These outfits with one-time arrangement relapse line for each gathering. It can, in the future, be amplified to cover the time arrangement information, with more computation control; it can be carried out at an hourly premise. This gives a more drilled/scrutinized investigation of the time-series data. [9]

This investigation survey gives a comparative time arrangement examination of the ground level and the remotely detected method of discussing quality estimation, appearing the outcomes in the best accessible determination to date and the day-by-day variety of the poisons and discussing quality file in different plots. This arrangement of the considerate too included the determination of molecule matter 2.5. Concurring with the think, the discussed quality made strides by and large amid the lockdown periods, but when the limits were released, the sum of poisons expanded once more. It recommends way better quality of discussion in the year 2020 as compared to 2019 and 2021. 2.64% for the year 2020 and 5.27% for the year 2021, and an expanding rate of 3% was seen after the lockdown period. For the years 2020 and 2021, the levels of all the criteria toxins were less as compared to the year 2019 and too had a diminishing slant for the lockdown stages but for Ozone, which has an expanded concentration in the lockdown period. O3 appeared an expanding drift for the year 2020, the reason owing to the barometrical chemistry between NO2 and O3. R2 values of the GWR demonstrate for the progressive, a long time are 0. [10]

The investigation thinks about surveying the execution of the three best information mining models for foreseeing the precise AQI information in a few of India's most crowded and contaminated cities. The manufactured minority oversampling method was utilized to equalize the course information to get superior and steady results. It was fruitful to accomplish higher exactness by utilizing this novel strategy of adjusting the datasets, utilizing them, and at that point closely comparing the outcomes of both imbalanced and adjusted ones to guarantee that the adjusted ones were profoundly exact. Measurable methods like RMSE, MAE, MSE, and R-SQUARE were at that point utilized to affirm the superior outcomes. The calculations were run utilizing both datasets (with and without the Destroyed calculation), and an increment of 6 to 24% was found- 6% in the same city and calculation. In the proposed work, irregular timberland relapse and CatBoost relapse reliably delivered empowering outcomes after exhaustive testing of all three calculations was conducted in Modern Delhi, Bangalore, Kolkata, and Hyderabad. It shows that the CatBoost and irregular timberland calculations, when combined with destroyed utilized datasets, can abdicate amazing results for AQI utilize cases in India. This, in turn, may empower neighbourhoods, states, and governments, along with other civic specialists, to take activity and direct the quality of the discussion. As it is clear from the measurements over, utilizing these relapse models on the AQI information from 2015 to 2020 has been effective in demonstrating that our inventive utilisation of the destroyed calculation has paid off and expanded the precision values of these relapse models. The objective is to distinguish patterns and offer recommendations for raising a city's discuss quality score. It is beneficial to examine the primary contributing causes and viable approaches to decrease them. [11]

The survey deals with dissecting and anticipating discussion quality utilizing information from two particular zones in Kolkata- Victoria and Rabindra. With its accentuation on hyper-parameter tuning to achieve the best exactness, this is abnormal in that it gives smart data around the adequacy of machine learning calculations for discussing quality expectations. Because of the energetic environment, erratic nature, and an assortment of toxins input and time, anticipating discussion quality can be troublesome. The negative results of discussing contamination on individuals, creatures, plants, chronicled places, the climate, and the biological system make reliable discussion quality fundamental. There has been a careful examination of the discussed contamination at two areas in Kolkata, to be specific Victoria and Rabindra, which contain six major discussed toxins, counting PM10, and PM2.5. The examination and examination of the information included information normalization, clear esteem filling, copy expulsion, and exception end. The best exactness of 97.98% was found on the Rabindra dataset utilising the SVC show, whereas the most elevated exactness of 93.29% was found from the Victoria dataset utilising the arbitrary woodland show after performing 5 fundamental classification calculations with suitable hyperparameter tuning. In the future, a dataset with a longer time period can be utilized to test an outfit machine learning classification calculation, which will empower the consolidation of meteorological components like wind speed and temperature in the AQI course expectation. [12]

The research thesis examines the efficacy of air pollution on public health, particularly the effects period of the pre and post covid-19 eruption and it also shows the needfulness the regulate the quality of air, especially in developing nations and associated with contagious diseases i.e. including Lungs Cancer, Autism, Asthma, and Low Birth Weight. In this review, the uses of a machine-learning algorithm are to analyse the concentration of notable air pollutants like SO2, NO2, PM2.5, O3, and PM10. This research identified major reductions in pollutants such as PM2.5, PM10 and NO2. In the time of social distancing and comparing the AQI data of pre and post covid-19 pandemic, which shows that human activities are a major control for the quality of air which implies which environmental measures to be taken such as reducing industrial emissions and other factors like to decrease in traffic congestion, which can improve the air quality, this paper presents the machine learning model Linear regression and popular Time Series algorithms such as LSTM, ARIIMA, and Prophet and the experiment result shows us the efficiency of the model in detecting the air quality index and future prediction of pollutant level. Generally, this survey elaborates on the condition of human activity, pollution of air, and public health. In this context different techniques of machine learning provide the practical approach toward the prediction of future pollution levels and analysis of the air quality index, and accordingly, important measures to be taken to reduce public health and the environment largely. [13]

Globally air pollution has always been a large concern to all and due to its worst effects. Accurate prediction in forecasting the air quality and level of concentration of various ambient air pollutants are the most vital for efficacious control of pollution. This survey searches various techniques which are very useful for the prediction of AQI and forecasting future pollutant concentrations, understanding AQI is very crucial ever since it serves as an indicator of the level of air pollution and thus induces the awareness of people towards their fatal effects and also attempting to decrease it accordingly. This research reviews the methodologies and follows by emphasizing the Artificial Neural Network (ANN) and logistic Regression for predicting the AQI and hence pollutant level prediction. This review concludes that the study in future on this scope includes other parameters like meteorological data with other pollutant concentrations for the AQI (Air Quality Index) and to forecast more precisely. [14]

The increasing severity of air pollution in the world, including India, is of concern due to its negative impacts on human health. Prediction and forecasting of the levels of air quality and various ambient air pollutants are crucial in taking effective pollution control measures. Therefore, the literature review tries to find the techniques used for forecasting AQI and the levels of pollutant concentrations in the future. Understanding AQI is essential because it acts as an index of the status of air pollution, creating awareness among the public of its negative impacts and the need to reduce it. The review goes through the methodologies which the researchers used to elaborate on techniques such as Artificial Neural Network, Linear Regression, and Logistic Regression for prediction of AQI and pollutant concentration forecasting. The review of the literature shows that most of the researchers have focused their work on AQI and pollutant concentration forecasting, using a different kind of modelling approach. ANN, Linear Regression, and Logistic Regression are also other techniques common in forecasting AQI and pollutant levels. The review suggests that future studies could widen their scope by incorporating additional parameters such as meteorological data along with pollutant concentrations for the prediction and forecasting of AQI. In essence, our literature review aims to provide a deep insight into AQI prediction and pollutant forecasting techniques and emphasizes the importance of such an attempt to address this ongoing crisis. [15]

The concern of air pollution is escalating and there is a demanding need for effective monitoring systems to protect human health and ecological balance. It explains the effect of poisonous emissions from industries and vehicles on the quality of air, which further presents health hazards. The paper suggests the integration of ML technology into air pollution monitoring systems to predict the level of pollution in the future based on historical data. It specifies how a device can be implemented, collecting real-time pollution data, processing that data using ML algorithms, and storing the results for further analysis. The study comes with forecasting various pollutants such as CO, SO2, O3, NO2, PM2.5, and PM10, using the data on meteorological parameters to improve accuracy in air quality forecasting and Introduction to Air Pollution Monitoring: This paper gives importance to continuous monitoring of air pollution due to its harmful effects on human health and the environment. This paper gives importance to determining the sources, intensity, and origins of pollution to take effective control measures. Utilization of Machine Learning: This paper suggests using ML algorithms for the estimation of the future level of pollution. Several ML techniques, including Linear Regression, Decision Tree, Random Forest, and Artificial Neural Networks, are used to predict air pollution.

Prediction Accuracy: The paper publishes the results of the prediction accuracy for different ML algorithms for various pollutants. Results indicate that the Random Forest algorithm yields better prediction accuracy for the Air Quality Index (AQI) compared to other methods. Experimental Setup: The study outlines an experimental setup of data collection from pollution monitoring stations, its software implementation through Anaconda Python, and the methodology of how machine learning modelling will be used. The paper concludes that ML-based prediction models for air quality give promising results. Suggestions are made for further research for the improvement of prediction accuracy and to increase the scope to include additional pollutants and metrological parameters. Overall, the paper provides insight into leveraging ML techniques for air pollution prediction and points out the importance of continuous monitoring for mitigating the devastating effects of air pollution on human health and the environment. [16]

The study aimed to predict air pollution levels (PM2.5, O3, NO2) and aerosol optical depth (AOD) at an urban traffic location in New Delhi using a modelling approach time series simulation. Methodologically, satellite aerosol data (AOD550) from Terra MODIS (Collection 6) were analysed, as well as average monthly concentrations of air pollutants and AOD from five 2012 to 2017. Simulations were performed with PM2.5, O3, NO2, and AOD over the same period, with forecasts extending to 2020-2023. Validation was performed on 24 months of in situ and satellite data from 2018 to 2019. Results show discrepancies in 2020 due to pandemic-related shutdowns, consensus between modelled/forecasted data and observed data, and the effectiveness of the autoregressive integrated moving average over series time like ARIMA to simulate and predict air pollution. Data taken from CSIR-Central Road Research Institute (CRRI) and Central Pollution Control Board (CPCB) India for air pollutant concentration (2012-2020) and MODIS AOD data for CSIR station -CRRI (2012-2020). The analysis shows an asymmetric and leptokurtic distribution of pollutants from 2012 to 2017, a significant decrease in pollutant concentrations in 2020 due to the lockdown, and the fit of ARIMA to the MODIS AOD forecast, PM2.5 and NO2 in New Delhi city. The study shows that ARIMA, combined with satellite data, can be a useful tool for predicting aerosols in areas with limited ground-based data. The study ended with the following model accuracy values:

- In the case of PM2.5 it is a good fit (Stationary R-squared: 0.752, R-squared: 0.648), with moderate errors (RMSE: 32.084, MAPE: 26.522%), but predictors are not statistically significant (p-value: 0.1025).

- For NO2 it is a Moderate fit (Stationary R-squared: 0.585; R-squared: 0.383), moderate errors (RMSE: 17.672, MAPE: 41.479%), and statistically significant predictors (p-value: 0.0000).

- For O3 it is a good fit (stationary R-squared: 0.700, R-squared: 0.537), with moderate errors (RMSE: 16.246, MAPE: 23.850%), and statistically significant predictors (p-value: 0.0001).

- In the case of AOD_{550}: Excellent fit (Stationary R-squared: 0.744, R-squared: 0.824), with low errors (RMSE: 0.105, MAPE: 9.236%), but predictors are not statistically significant (p-value: 0.1076). [17]

After studying the paper, we got to know that this paper is aimed to predict air pollutant concentration in major cities of India like Delhi and Agra. The need for this paper arose due to the rising level of various air pollutants which in turn is causing various diseases. The authors collected a comprehensive dataset of various air pollutants, weather and traffic data from sources like the Central Pollution Control Board of India over a specific period of time. The dataset is fitted to an LSTM model with input of weather data, traffic conditions, festivals and holidays to accurately predict air pollutant levels. The model is then tested with some pre-existing data. The predicted values are compared with the original values to determine the model's effectiveness. To calculate the effectiveness the authors used Root Mean Squared Error (RMSE) values. The calculated RMSE values were less than 15 for 12-hour forecasts, less than 8 for 6-hour forecasts and 5 for 1-hour forecasts. Going to conclusion the paper emphasizes the accuracy of air pollution forecasting. It also recommends additional research to predict air pollution levels of other cities of India with the help of sensors, which in turn may help the local governmental bodies to regulate policies to minimize air pollution. [18]

After studying the paper on predicting air quality and pollution levels in different regions we get to know it uses statistical methods and machine learning techniques to predict and forecast the same. The paper shows us the approach that combines statistical techniques like time series analysis and regression with machine learning techniques like Ada-boosting to predict PM 2.5 levels in Hyderabad. The authors have collected data from Kaggle. Then the dataset is fitted to model Linear Regression, Ada-Boosting and XG-Boosting. The models are then used to predict PM 2.5 values based on different atmospheric and surface parameters. The values are then compared with the original values to calculate effectiveness via parameters like RMSE, MAE and R2_score. In the case of linear Regression, R2_score is around 0.5 and for Ada-Boosting, R2_score is around 0.38. In conclusion, air pollution is a significant environmental issue, and improving air quality by forecasting it can have a positive impact on public health. The paper also suggests that future work could be forecasting using deep learning techniques to improve accuracy. [19]

This paper focuses on the importance of raising awareness about air pollution in urban areas and presents an IoT-based system designed to monitor and forecast air pollution over real-time data, with a special focus on the city of Skopje. It emphasizes that a significant proportion of the world's population lives in areas where the air quality exceeds WHO guidelines, leading to millions of deaths each year. The capital of North Macedonia Skopje has a severe problem with air pollution due to exposure to PM2. 5 which results in early mortality and high social costs. The existing Skopje air quality monitoring system has several drawbacks such as poor pollutant measurements low spatial resolution and malfunctioning sensors. The study suggests an IoT-based system that attempts to increase the value of data by creating an appropriate data processing subsystem to throw light on these issues. This subsystem focuses on enhancing the prediction of future pollution levels and the amount of time required to reach alarm thresholds by utilizing cutting-edge machine learning techniques particularly deep learning. The data processing subsystem includes several key steps, such as outlier analysis, data smoothing, handling missing records, and merging weather and sensor data. Deep learning techniques, such as principal component analysis and deep belief networks, are used to augment data and improve the reliability of predictive models. The pollution forecasting model focuses on the pollutant PM10 and uses a bidirectional short- and long-term memory model (BiLSTM), the performance of which is evaluated using the mean squared error original cylinder (RMSE) on the PM10 Pollutant Test Set in the central city of Skopje.

MSE values of various pollutants such as CO are 0.21, for NO2 is 0.06, for O3 26.13, for PM 10 0.22, for PM 2.5 it is 7.32 and SO2 it is 0.08. Although the resolution of the data is low, the results of the pollution prediction model are promising, thus demonstrating the potential effectiveness of the proposed system.

In conclusion, this paper integrates this module with IoT infrastructure for air pollution detection which will significantly improve overall performance, further allowing for more informed decisions and implementation of steps to proactively lower urban air pollution. [20]

2.1 Tabular Observation of Literature Review

Serial Number	Title	Observation
1	Time series analysis of the Air Quality Index for several Indian cities [1]	The research uses VARMA, LSTM and VAR models to predict AQI. It uses MSE as a performance metric.
2	A Deep Learning Approach for Forecasting Air Pollution in South Korea Using LSTM [2]	The research uses the Encoder-Decoder (En-De) model and LSTM model to predict air pollution. It uses performance metrics like RMSE, MSE and MAE
3	Time Series Analysis for Air Quality Forecasting [3]	The research uses ARIMA and LSTM to predict various air pollutants like SO2, NO2 and PM10 in two cities Margao and Sanguen. It uses performance metrics RMSE and MAPE.
4	Air Quality Index Prediction using LSTM [4]	The research uses VARMA, LSTM and VAR models to predict AQI. It uses MSE as a performance metric.
5	Air quality monitoring system based on IoT using Raspberry Pi [5]	The research is about how to monitor air pollutants (PM2.5, CO2and CO), temperature, relative humidity and pressure using Raspberry Pi.
6	Development of an IoT-based indoor air quality monitoring platform. [6]	The research is about how to monitor Particulate matter (PM2.55 and PM10), hydrocarbon-based petroleum products (CO, CO2) in the air and humidity using a microcontroller.
7	An IoT-based sensing system for modelling and forecasting urban air quality [7]	The research is about how to monitor air pollutants (PM2.5, CO2, CH4 and CO), temperature, relative humidity and noise using Arduino Uno. Then using the collected data, it predicts AQI using the ARIMA model. It uses performance metrics RMSE, MSE, MPE, MAPE and MAD
8	Air quality system using IoT for indoor environmental monitoring [8]	The system in the research measures indoor temperature, pressure, and air quality and makes the user aware of any undesirable change in the environment leading to problems in their lives
9	Air pollution prediction in smart cities by using machine learning techniques [9]	The research uses ETD, LSTM and ARIMA models to predict AQI.
10	Time Series Analysis of Air Quality of an Industrial District of India Using Remote Sensing and GIS: Impact of Covid-19 Lockdown [10]	The research is on the study of different air pollutants and how their concentration reached during a lockdown
11	Prediction of air quality index using machine learning techniques: a comparative analysis [11]	The research uses Support Vector Regression, Random Forest Regression and CatBoost Regression model to predict AQI. It uses MSE, RMSE, MAE and R-Square as performance metrics.
12	Air Quality Monitoring Using Statistical Learning Models for Sustainable Environment [12]	The research uses ARIMA, LSTM and Prophet model to predict AQI. It uses Precision, Recall, F1-Score and Support as performance metrics.
13	ANALYSIS AND PREDICTION OF AIR QUALITY INDEX IN INDIA DURING PRE AND POST COVID PANDEMIC USING MACHINE LEARNING ALGORITHMS [13]	The research uses ARIMA, LSTM and Prophet model to predict AQI. It uses R-Square as a performance metric.
14	A literature review on prediction of air	This research reviews the methodologies and

	quality index and forecasting ambient air pollutants using machine learning algorithms [14]	follows by emphasizing the Artificial Neural Network (ANN) and logistic Regression for predicting the AQI and hence pollutant level prediction
15	Air pollution prediction system for smart city using data mining technique: a survey [15]	This research reviews the methodologies and follows by emphasizing the Artificial Neural Network (ANN), Linear Regression, and Logistic Regression for predicting the AQI
16	Air quality prediction by machine learning [16]	The research uses Linear regression, Decision Tree and Random Forest regression models to predict various air pollutants.
17	Time Series Simulation and Forecasting of Air Quality Using In-situ and Satellite-Based Observations Over an Urban Region [17]	The research uses the ARIMA model to predict various air pollutants like PM2.5, NO2, O3 and AOD550. It uses RMSE, MAPE and R-Square as performance metrics.
18	Time series-based LSTM model to predict air pollutant concentration for prominent cities in India [18]	The research uses the LSTM model to predict AQI. It uses RMSE as a performance metric.
19	Prediction of Air Quality and Pollution using Statistical Methods and Machine Learning Techniques [19]	The research uses Linear Regression, Ada-Boosting and XG-Boosting models to predict PM2.5. It uses RMSE, MAE and R-Square as performance metrics.
20	Smart city air pollution monitoring and prediction: A case study of Skopje [20]	The research uses the BiLSTM model to predict PM10. It uses MSE as a performance metric.

3. System Overview and Design

Data mining is very important for uncovering patterns in large datasets, often revealing correlations that were previously unrecognized or ignored.

In this project several Deep Learning and Machine Learning algorithms were employed like ARIMA, SARIMA, PROPHET, LSTM, Liner Regression and Polynomial Regression, to analyse the best model fit for future forecasting.

3.1 Algorithms

We used various machine learning algorithms to evaluate our dataset. The machine learning models used are listed below.

ARIMA

ARIMA is an autoregressive integrated moving average. ARIMA utilizes time series data to analyze the dataset and forecast future trends. The implicit premise of autoregressive models is that the future will be like the past. It comprises autoregression (AR), differencing (I), and moving average (MA) components. The AR component involves regressing the variable on its past values, with the order (p) determining how many lagged values are included. Differencing (d) is used to make the series stationary, ensuring consistent mean, variance, and autocorrelation over time. The model involves steps such as identification (using ACF and PACF plots to determine p d q) estimation (fitting the model and estimating parameters) diagnostic checking (ensuring residuals resemble white noise) and forecasting (predicting future values with confidence intervals). The MA component models the error term as a combination of past error terms with the order (q) indicating how many lagged errors are used. The model is typically denoted as ARIMA (p d q). ARIMA's versatility enables it to manage types of time series data making it applicable, in fields such as economics, finance, environmental science, engineering and public health. For instance, in our air quality initiative, we utilize ARIMA to examine information obtained from the CPCB and IoT gadgets to predict levels of pollutants such, as PM 2.5 or PM 10.

SARIMA

SARIMA, short, for Seasonal Auto-Regressive Integrated Moving Average serves as a version of ARIMA by including aspects, alongside non-seasonal elements. This model is tailored to address recurring patterns effectively and helps in capturing both prolonged data relationships to enhance its capabilities.

Prophet

Prophet, a tool developed by Facebook is commonly used for predicting time series data by utilizing a model that accommodates trends through yearly, weekly and daily seasonal patterns. It proves effective when dealing with data showing variations and possessing multiple seasons of historical information. The tool demonstrates resilience in handling missing data points and outliers. Moreover, it is structured to factor, in holiday impacts and transition points enabling it to capture both short-term fluctuations and long-term tendencies. Its intuitive additive model comprises elements for trend analysis, seasonality assessment and holiday considerations making it versatile for applications. With its adaptability and reliability Prophet serves as a resource for forecasting across sectors such, as business, economic and environmental monitoring where datasets may exhibit intricate seasonal patterns and occasional disruptions.

Long Short-Term Memory

LSTM, known as Long Short-Term Memory belongs to a type of network (RNN) designed specifically to understand and learn extended relationships, in sequential data. Unlike RNNs LSTM networks have a structure that involves memory cells and gating mechanisms like input, output and forget gates. These gates play a role in controlling the information flow within the network enabling it to retain details over long periods while discarding irrelevant data. This capability to grasp long-term dependencies makes LSTM particularly effective for predicting time series data in datasets with temporal patterns. It finds applications across domains such as Natural Language Processing, Speech recognition and environmental monitoring due to its proficiency in handling sequential data with temporal links and providing precise forecasts, for future occurrences.

Linear Regression

Linear Regression is a technique that is commonly used to establish the connection, between a dependent variable and one or more independent variables by creating a linear equation based on collected data. With X representing the independent variable Y, the dependent variable $b1$ the slope of the line and $b0$ the y-intercepts a straight line is plotted using the equation $Y=b0+b1X$. $A=B_0+C_1+D_2+C+. +b_nX_n$ which is expanded and includes variables in Multiple Linear Regression. The primary aim of Linear Regression is to minimize the sum of variances between observed and predicted values. It finds applications in fields such, as finance, engineering and environmental science by aiding in grasping underlying patterns and making forecasts based on past data.

Polynomial Regression

Polynomial regression establishes a relationship between the independent variable (X) and the dependent variable (Y) which is represented as an n-degree polynomial. More complex relationships are captured by the Polynomial regression that fits in a curved line. The polynomial regression equation used generally is $Y=b0+b1X+b2X^2+b3X^3+...+bnX^n$. Here, b0, b1, b2, b3... bn are the formula's special numbers, and n tells you how curvy the line is. This curve line helps us see more complicated connections between X and Y, especially when their relationship isn't straight but rather curves up or down. By making the curve line more curvier (increasing 'n'), it can follow the ups and downs of the data more closely. But, if we make it too curvy, it might start fitting too closely and not be useful for understanding the bigger picture. In applications like environmental science, polynomial regression can effectively model phenomena such as pollutant concentration levels, which may exhibit nonlinear trends over time.

3.2 System Workflow

We used various machine learning algorithms to evaluate our dataset. The machine learning models used are listed below.

The following diagrams illustrate how our machine-learning model operates and how the IoT device collects data and transmits it to our database for future use.

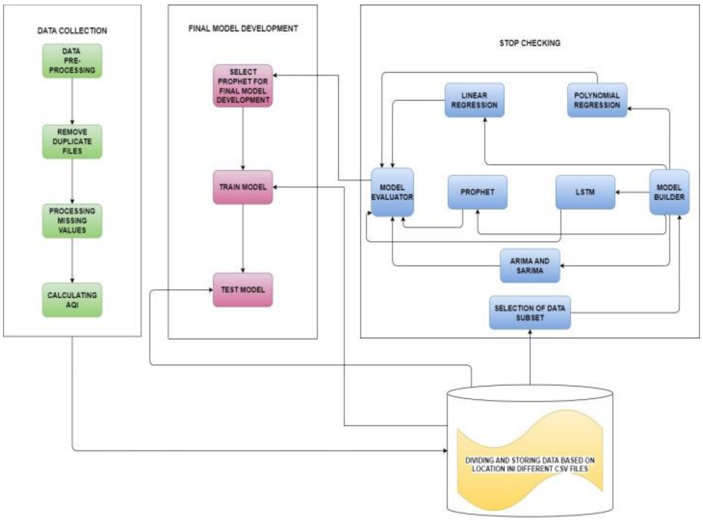


Figure 1: Proposed Workflow of The Model

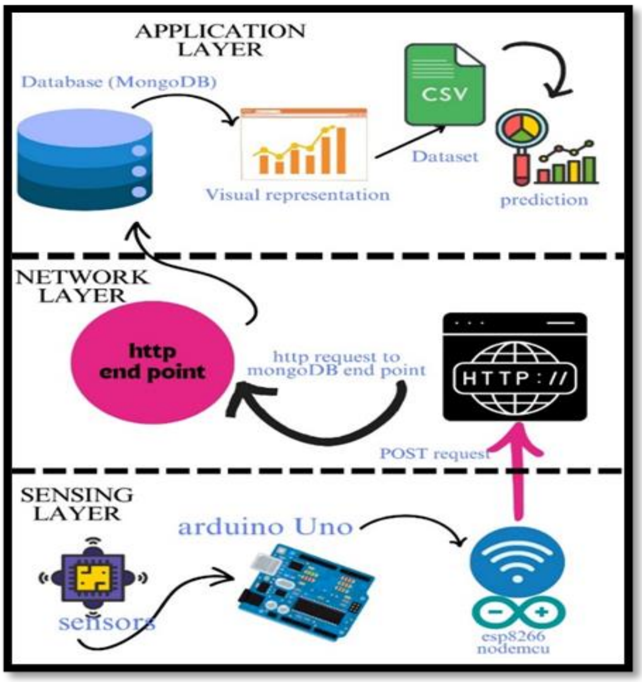


Figure 2: Proposed Workflow Of The Hardware Device

4. HARDWARE SETUP

To collect data on various air pollutants, we developed an IoT device. The structure and functioning of the IoT device are depicted below.

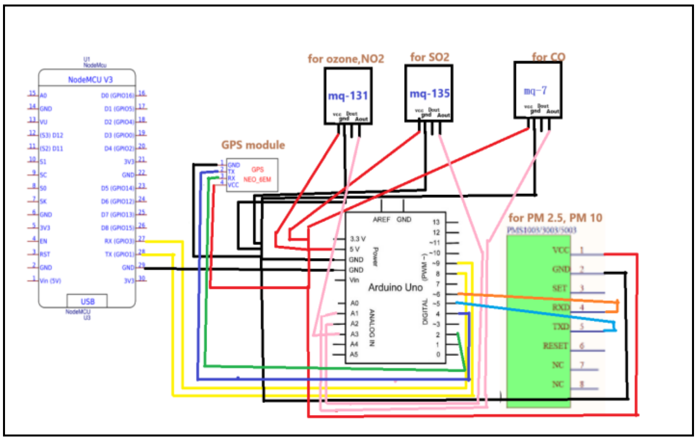


Figure 3: Circuit Diagram of Hardware Setup

4.1 Components and their Usage in IoT Devices

In our setup, we have used **Arduino Uno** as it has 6 analogue pins to take input from more than one sensor. It also has 14 digital pins which are further used as Software Serial for data transfer. In sensors, we used **MQ-7, MQ-135**, and **MQ131** gas sensors. **MQ7** is the carbon monoxide sensor. **MQ135** has been used to detect Sulphur dioxide and Nitrogen Dioxide and lastly, **MQ131** as the ozone gas sensor. All the sensors have a VCC-GND, AOUT and DOUT depictive of the +5v input voltage required, analogue output and digital output respectively. The Tin Oxide (SnO2) is the main sensing element in the above-stated sensors. As per the analogue output from these, the MQUnified gas sensors package in the Arduino library is used to calibrate them and detect the gases.

PMS5003 digital infrared dust sensor is used to get the correct concentration of the suspended Particulate matter PM 2.5 and PM 10 in the air.

Esp8266 Nodemcu is used for the WIFI module to transfer data. The RX/TX pins of it are alternatively connected with the Arduino Uno.

Neo-6M GPS is used to get the latitudinal and longitudinal location of the measuring station.

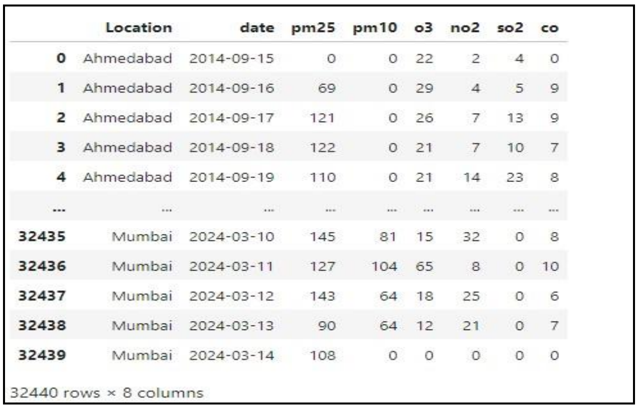
4.2 Working

The IoT model is focused on gathering concentrations of various air pollutants, namely Carbon monoxide, Ozone, Sulphur dioxide, Nitrogen dioxide, and particulate matter - PM 2.5 and PM 10 by utilising MQ series sensors and the PMS5003 sensor for this purpose. MQ-131 detects ozone gas, and MQ-135 detects SO_2 and NO_2 gases. MQ-7 detects Carbon monoxide. PMS5003 shows PM 2.5 and PM 10 concentrations in the air. Data is collected using an Arduino UNO board where all the Vcc are connected to 5V and GND to GND. Analog data received in the sensors are converted to digital reading using Arduino packages of the MQUnified sensors package library and PMS5003 Package library. The NEO-6M GPS is used to detect the latitude and longitude of the centre. The data is then converted into a concatenated string and passed to Esp8266 Nodemcu. The Nodemcu connections involve alternate Rx and Tx connections for data transfer. It is connected to a specific WIFI node. The Nodemcu data is subsequently transferred via HTTP request in the HTTP body to the HTTP End point of the dedicated database hosted on a MongoDB server. This data collection is specifically tailored for indoor air quality measurement. This dataset will serve as a foundation for subsequent efforts in predicting indoor air quality.

5. DATA COLLECTION AND PREPROCESSING

The data collection was carried out in two parts: one involving data from the CPCB website and the other from our IoT device.

CPCB Data: We collected the main data from the Central Pollution Control Board (CPCB) of India (official website: https://cpcb.nic.in), which is a statutory organization under the Indian government. The dataset that we collected had 32440 entries with 8 fields (Location, Date, PM25, PM10, O_3, NO_2, SO_2, CO). The size of the dataset is around 2.0 MB. The sample of the dataset is below.



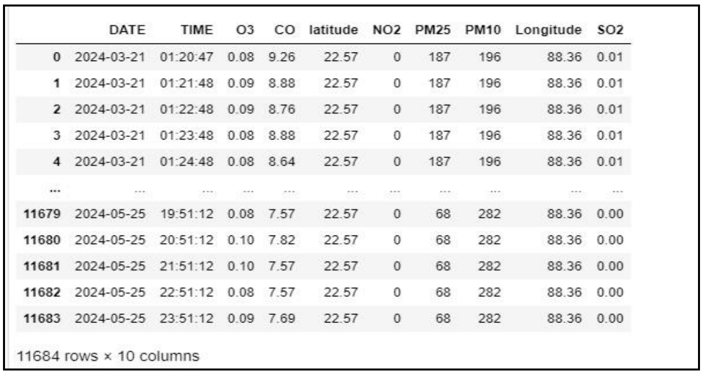
	Location	date	pm25	pm10	o3	no2	so2	co
0	Ahmedabad	2014-09-15	0	0	22	2	4	0
1	Ahmedabad	2014-09-16	69	0	29	4	5	9
2	Ahmedabad	2014-09-17	121	0	26	7	13	9
3	Ahmedabad	2014-09-18	122	0	21	7	10	7
4	Ahmedabad	2014-09-19	110	0	21	14	23	8
...	...	...	...	...	...	...	...	...
32435	Mumbai	2024-03-10	145	81	15	32	0	8
32436	Mumbai	2024-03-11	127	104	65	8	0	10
32437	Mumbai	2024-03-12	143	64	18	25	0	6
32438	Mumbai	2024-03-13	90	64	12	21	0	7
32439	Mumbai	2024-03-14	108	0	0	0	0	0

32440 rows × 8 columns

Figure 4: Air pollutant dataset in India (2014 - 2024) collected from the CPCB website

IoT Device Data: Data was collected daily from an IoT-based device developed as part of the project, which was located in Kolkata. This device continuously monitors and records air quality parameters in real-time. The data collection was done on an hourly basis and the current CSV file contains 11684 rows and 10 columns, and the size of data is 912.9 KB. The dataset collected using the device has been uploaded to the Kaggle website.

The link to the website: https://kaggle.com/datasets/794d435b73dabe0a762cd5f6b93911ef1f058bf8d25f24bcd0a8971306f02fa2



	DATE	TIME	O3	CO	latitude	NO2	PM25	PM10	Longitude	SO2
0	2024-03-21	01:20:47	0.08	9.26	22.57	0	187	196	88.36	0.01
1	2024-03-21	01:21:48	0.09	8.88	22.57	0	187	196	88.36	0.01
2	2024-03-21	01:22:48	0.09	8.76	22.57	0	187	196	88.36	0.01
3	2024-03-21	01:23:48	0.08	8.88	22.57	0	187	196	88.36	0.01
4	2024-03-21	01:24:48	0.08	8.64	22.57	0	187	196	88.36	0.01
...	...	...	...	...	...	...	...	...	...	...
11679	2024-05-25	19:51:12	0.08	7.57	22.57	0	68	282	88.36	0.00
11680	2024-05-25	20:51:12	0.10	7.82	22.57	0	68	282	88.36	0.00
11681	2024-05-25	21:51:12	0.10	7.57	22.57	0	68	282	88.36	0.00
11682	2024-05-25	22:51:12	0.08	7.57	22.57	0	68	282	88.36	0.00
11683	2024-05-25	23:51:12	0.09	7.69	22.57	0	68	282	88.36	0.00

11684 rows × 10 columns

Figure 5: Data collection done by IOT device (March 2024 - May 2024)

5.1 DATA PREPROCESSING

When we loaded the data, we found the data was not suitable for our work. It had problems with its type and had many missing values. So, the data had to be processed in a useful format.

Data processing was done in the following steps:

i. Data Cleaning

ii. Data Transformation

iii. Data Reduction

Data Cleaning: This process involves removing unwanted data like duplicate data, unformatted data and incorrect data.

- In our dataset we saw that the format of the data was all object so we had to convert them to their respective type.
    - The date field had to be converted to a date-time format.
    - The location field had to be converted to string format.
    - The various air pollutant fields had to be converted to integer format.
- Then we found many data had similar dates and locations so we removed the copy.
- As for checking for incorrect data we did not do it, as we took the data from a government organization.
- As for the missing values we filled it with the nearest data because usually pollutants do not change much in the next day.

Data Transformation: This process involves the improvement of the structure of data. This helps in better data-driven decision-making. This involves Normalization, Attribute Selection, Discretization and Concept Hierarchy Generation.

- We made an AQI field for our usage.
- We normalized data before fitting it to various models.
- We converted the daily data to a monthly mean value before fitting it to the SARIMA model.

Data Reduction: This process involves reducing the size of the dataset while preserving the important information. Some common steps involved are Feature Selection, Feature Extraction, Sampling, Clustering and Compression.

We divided the whole dataset by categorising it according to location

5.2 DATA PREPROCESSING FOR IOT DEVICE

IoT data device was used to collect pollutant data continuously on a daily basis. Each day, the Air Quality Index (AQI) was calculated from the collected data. The average AQI for each day was then determined. Upon reviewing the data, it was found that there are records for 53 consecutive days. This data collection period starts from March 21, 2024, and extends to May 25, 2024. During this entire period, the device consistently collected pollutants and AQI was calculated, ensuring a comprehensive set of daily average AQI values.

6. MODEL DEVELOPMENT

The data collection and data preprocessing were initially where duplicate values were removed and missing values were processed. Then eventually the respective AQI was calculated corresponding to a unique date. The data was further divided and stored based on locations in different CSV files.

6.1 DATA VISUALIZATION

When we loaded the data, we found the data was not suitable for our work. It had problems in its type and had many missing values. So, the data had to be processed in a useful format.

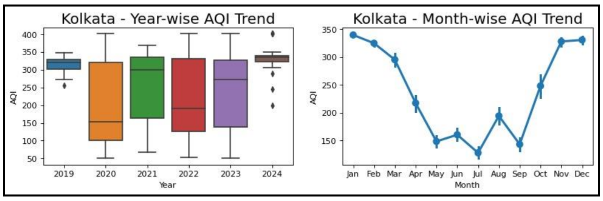


Figure 6: Data Visualisation of AQI for the city of Kolkata

6.2 IMPLEMENTATION USING ARIMA AND SARIMA MODEL

ARIMA and SARIMA were performed on the data to forecast AQI values for different cities. The SARIMA modelling process was defined to handle each city. The data was resampled to monthly averages and plotted to visualise AQI trends over time. Along with it, a boxplot was plotted to show AQI range distribution. The AQI time series was split into training and testing sets, and the best ARIMA model parameters were determined using the 'auto_ARIMA' function. A SARIMA model was fitted to the training data and predictions were made on the test set. The plotting of predictions against test data and the entire dataset was done including future predictions.

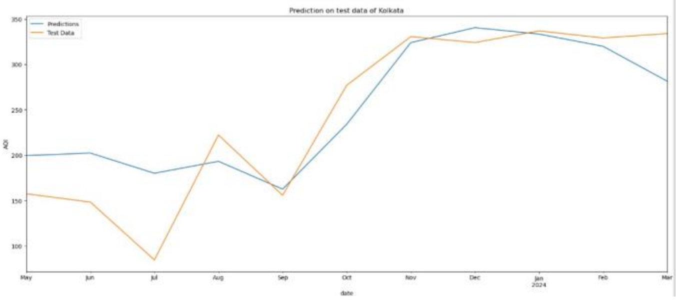


Figure 7: Graph between AQI and Date on Prediction and Test Data

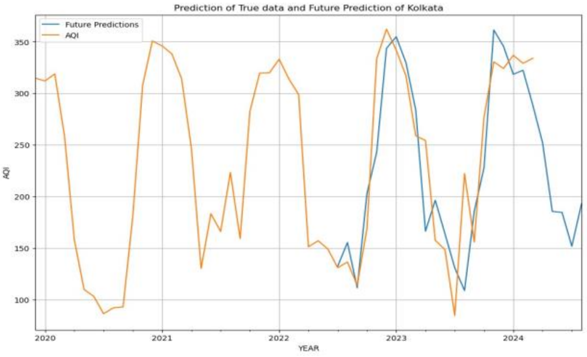


Figure 8: Graph between AQI and Date on Future Predictions and Original Data

6.3 IMPLEMENTATION USING LONG SHORT-TERM MEMORY

Long Short-Term Memory was used to forecast air quality data, that is AQI. The dataset was constructed and trained and the LSTM model was evaluated. The model is built with an LSTM layer and Dense layer, compiled using Adam optimizer and mean square error loss function. The model was fitted to training data and predictions were made. Both actual and predicted AQI values for the test set are plotted and also future AQI values were forecasted based on unique city names for a specified number of steps, visualizing these predictions alongside historical data.

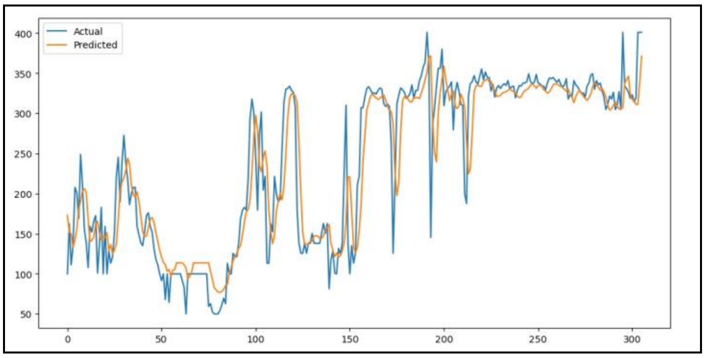


Figure 9: Graph between AQI and Date on Actual and Predicted Data

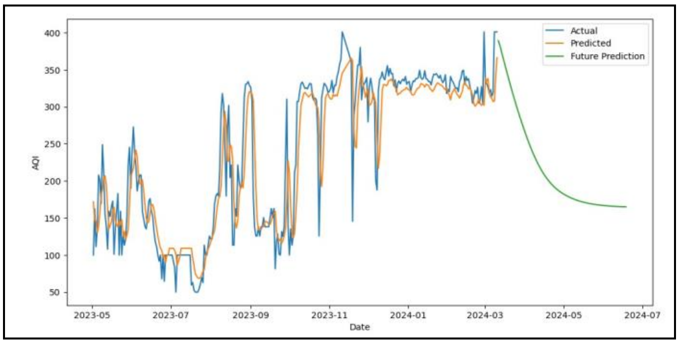


Figure 10: Graph between AQI and Date on Actual, Predicted and Future Prediction Data

6.4 IMPLEMENTATION USING PROPHET

The Prophet forecasting tool was used to predict the AQI for various cities. The data was processed and forecasting was performed. The creation of new columns was made 'ds' and 'y' which are required by Prophet from 'date' and 'AQI', respectively, and then data was resampled to monthly averages, filling in missing values. The data set was split into training and testing and then the function was initialized and the Prophet model was run on the dataset. Predictions were made on the test set. Then graph was plotted on the forecasted values against the actual values, with the city name displayed on the plot. It was then fitted to another Prophet model to the entire dataset and it also generated a future forecast for the next 200 days, thus visualizing these predictions along with actual values. In addition to the overall forecast, the trend was plotted, and yearly seasonality components were extracted by the Prophet model which provided insights into the underlying patterns in the AQI data. These visualizations helped us to understand both long-term trends and seasonal variations.

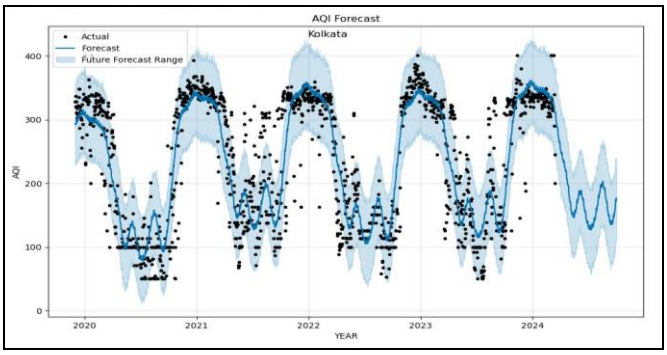


Figure 11: Graph between AQI and Date on Actual and Forecast Data with Future Forecast Range

6.5 IMPLEMENTATION USING LINEAR REGRESSION

Linear Regression was used to predict the AQI for different cities using historical data. The data was split into training and testing sets and then fitted to a linear regression model. The plotting was done for actual vs predicted AQI values and also future predictions of 200 days are generated.

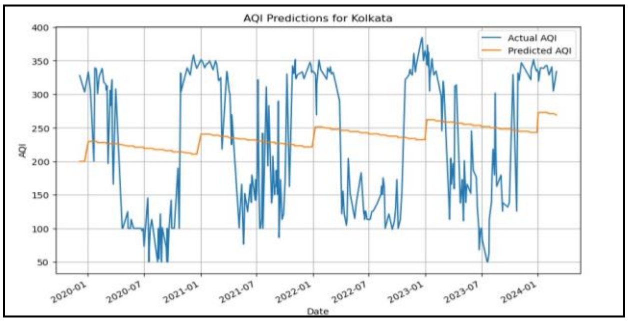


Figure 12: Graph between AQI and Date on Actual and Predicted Data

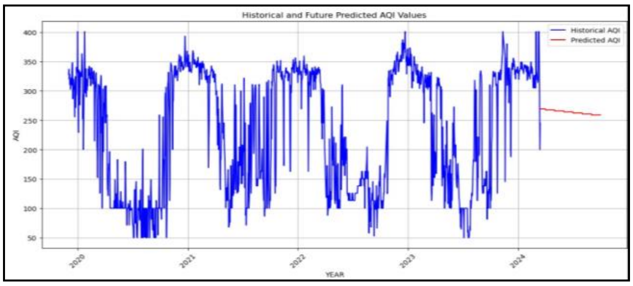


Figure 13: Graph between AQI and Date on Original and Future Predicted Data

6.6 IMPLEMENTATION USING POLYNOMIAL REGRESSION

Polynomial Regression was used to predict the AQI for various cities based on historical data. The data were processed and polynomial features were created and added for regression. The linear regression model was fitted to polynomial features and data were split into training and testing sets. It generated future AQI predictions and plotting was done.

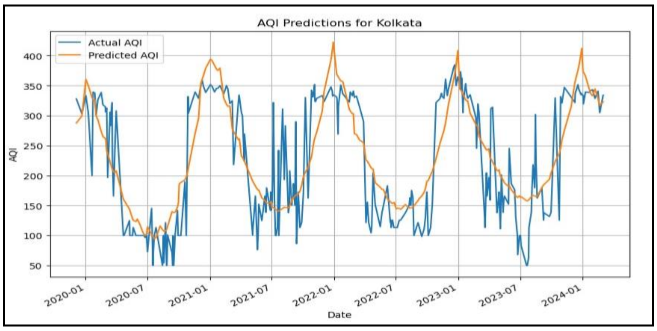


Figure 14: Graph between AQI and Date on Actual and Predicted Data

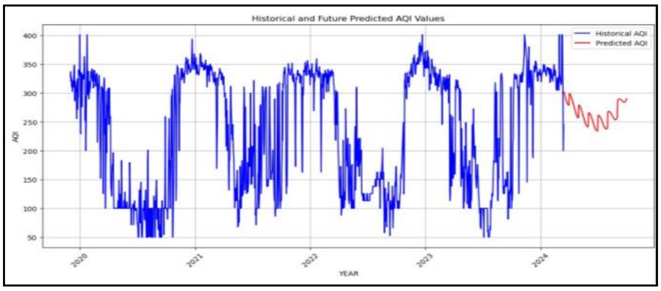
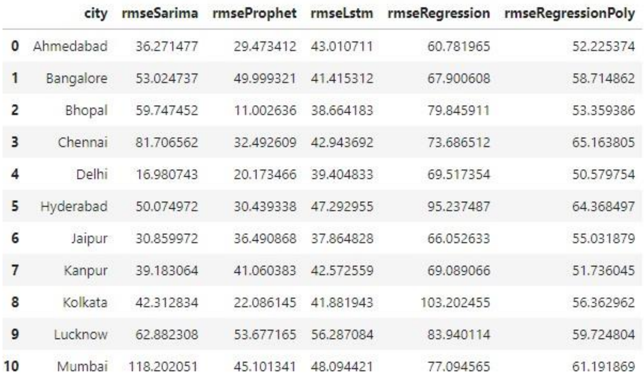


Figure 15: Graph between AQI and Date on Original and Predicted Data

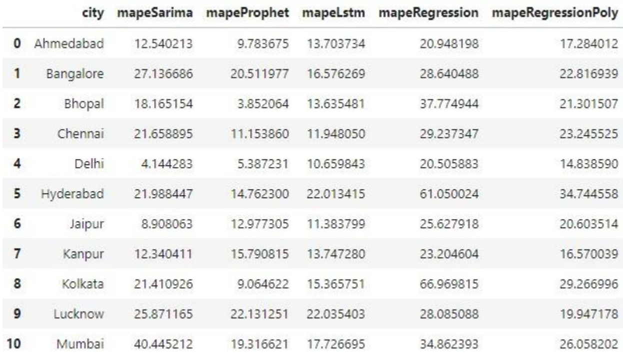
6.7 COMPARISONS OF MODELS FOR ALL LOCATIONS

The model implementation was done in all 11 cities. The cities are Ahmedabad, Bangalore, Bhopal, Chennai, Delhi, Hyderabad, Jaipur, Kanpur, Kolkata, Lucknow and Mumbai. Like Kolkata, as shown above, a similar evaluation was done for the other 10 cities as well.



	city	rmseSarima	rmseProphet	rmseLstm	rmseRegression	rmseRegressionPoly
0	Ahmedabad	36.271477	29.473412	43.010711	60.781965	52.225374
1	Bangalore	53.024737	49.999321	41.415312	67.900608	58.714862
2	Bhopal	59.747452	11.002636	38.664183	79.845911	53.359386
3	Chennai	81.706562	32.492609	42.943692	73.686512	65.163805
4	Delhi	16.980743	20.173466	39.404833	69.517354	50.579754
5	Hyderabad	50.074972	30.439338	47.292955	95.237487	64.368497
6	Jaipur	30.859972	36.490868	37.864828	66.052633	55.031879
7	Kanpur	39.183064	41.060383	42.572559	69.089066	51.736045
8	Kolkata	42.312834	22.086145	41.881943	103.202455	56.362962
9	Lucknow	62.882308	53.677165	56.287084	83.940114	59.724804
10	Mumbai	118.202051	45.101341	48.094421	77.094565	61.191869

Figure 16: RMSE values of various cities



	city	mapeSarima	mapeProphet	mapeLstm	mapeRegression	mapeRegressionPoly
0	Ahmedabad	12.540213	9.783675	13.703734	20.948198	17.284012
1	Bangalore	27.136686	20.511977	16.576269	28.640488	22.816939
2	Bhopal	18.165154	3.852064	13.635481	37.774944	21.301507
3	Chennai	21.658895	11.153860	11.948050	29.237347	23.245525
4	Delhi	4.144283	5.387231	10.659843	20.505883	14.838590
5	Hyderabad	21.988447	14.762300	22.013415	61.050024	34.744558
6	Jaipur	8.908063	12.977305	11.383799	25.627918	20.603514
7	Kanpur	12.340411	15.790815	13.747280	23.204604	16.570039
8	Kolkata	21.410926	9.064622	15.365751	66.969815	29.266996
9	Lucknow	25.871165	22.131251	22.035403	28.085088	19.947178
10	Mumbai	40.445212	19.316621	17.726695	34.862393	26.058202

Figure 17: MAPE values of various cities

6.8 MODEL IMPLEMENTATION FOR INDOOR AQI PREDICTION

DATA VISUALIZATION FOR INDOOR DATA

It was also observed that the AQI at night was more compared to day.

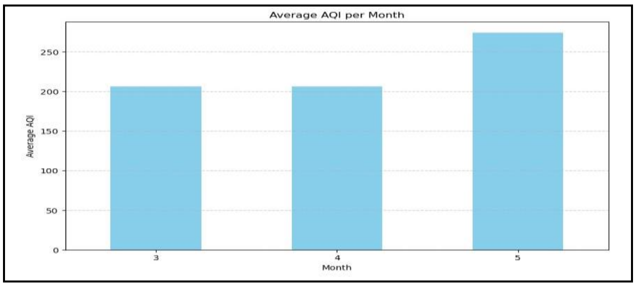


Figure 18: Monthly average AQI Visualization

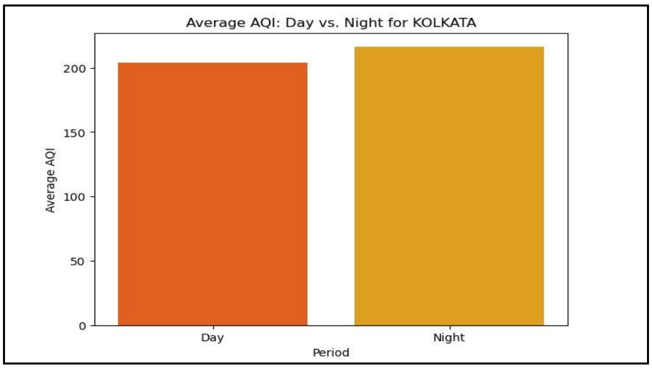


Figure 19: Average AQI during Day and Night

6.9 PROPHET MODEL IMPLEMENTATION FOR INDOOR AQI PREDICTION

After implementing the models, it was found that the best model evaluation was given by PROPHET. Thus, this model is also used in future predictions for indoor AQI.

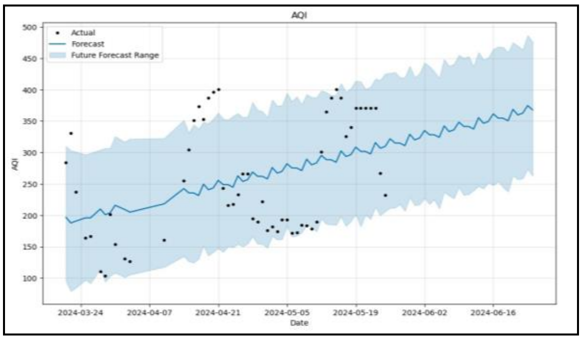
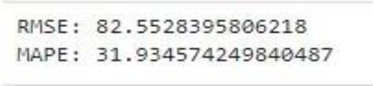


Figure 20: Graph between AQI and Date on Actual and Forecasted data with Future Range Data



```
RMSE: 82.5528395806218
MAPE: 31.934574249840487
```

Figure 21: RMSE and MAPE evaluation

7. MODEL PERFORMANCE COMPARISON

Performance evaluation is an important part of the implementation of the models. It helps us to understand the best model suitable for our project. The model performance has been carried out using two measures. The analysis is carried out based on computations which were done by comparing the actual values to the predicted outcomes. The Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) were used to check the performance of the model by performing necessary evaluations.

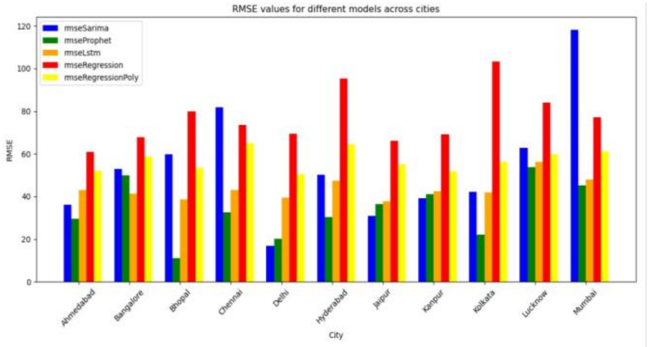
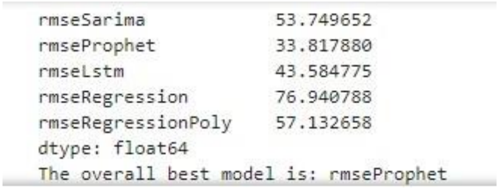


Figure 22: Performance Graph (RMSE)



```
rmseSarima             53.749652
rmseProphet            33.817880
rmseLstm               43.584775
rmseRegression         76.940788
rmseRegressionPoly     57.132658
dtype: float64
The overall best model is: rmseProphet
```

Figure 23: Best Model According to RMSE evaluation

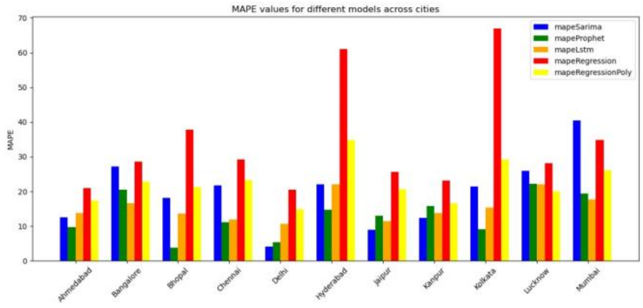
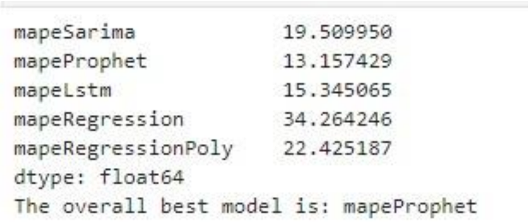


Figure 24: Performance Graph (MAPE)



```
mapeSarima             19.509950
mapeProphet            13.157429
mapeLstm               15.345065
mapeRegression         34.264246
mapeRegressionPoly     22.425187
dtype: float64
The overall best model is: mapeProphet
```

Figure 25: Best Model According to MAPE evaluation

After comparing it is found that the best model is Prophet. The model has a mean RMSE of 33.817880 and a mean MAPE of 13.157429 which is less than the other models, therefore implying that the Prophet model is good for forecasting AQI.

8. WEBSITE DEVELOPMENT

The website developed by our team offers a comprehensive air quality monitoring and prediction service. Users begin by registering their details in our MongoDB database. Once registered, users can log in to access the main dashboard. The dashboard features a search bar that utilizes API which calls to Waqi.info, allowing users to search for AQI and pollutant levels in specific cities and nearby monitoring stations. Additionally, users can enable GPS location services to detect their current location and display the AQI at the nearest monitoring station. The indoor air quality can also be monitored using our system which can display indoor AQI readings from nearby IoT sensors when GPS is enabled. This data is also retrieved from our MongoDB database.

A significant feature of our website is AQI prediction. Users can search by city and date to obtain AQI forecast for the next five days. This prediction covers both outdoor and indoor air quality and is powered by advanced machine-learning models developed specifically for our project.

The link to the website is: https://fnode-7lcl.onrender.com/

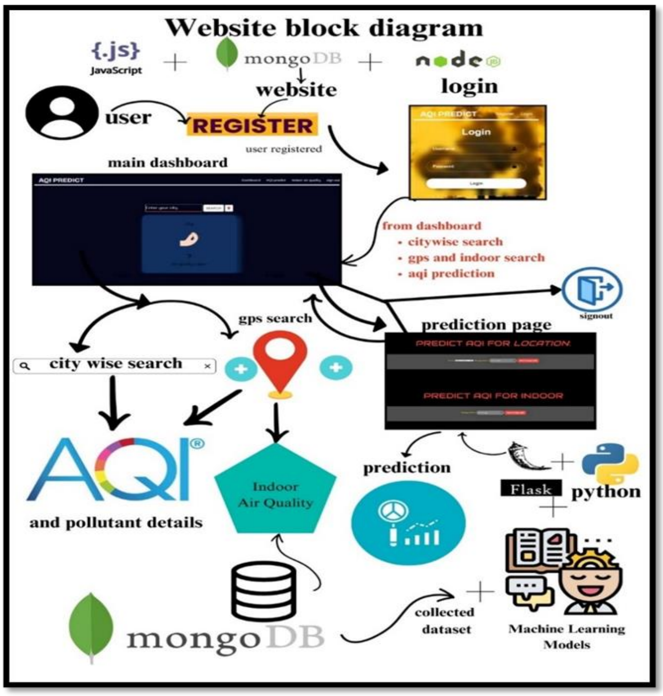
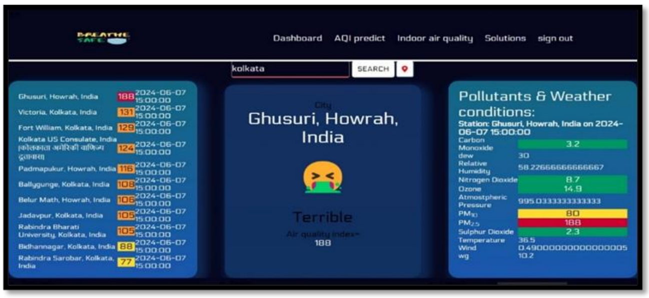


Figure 26: WORKFLOW DIAGRAM OF THE WEBSITE



Ghusuri, Howrah, India	188	2024-06-07 15:00:00
Victoria, Kolkata, India	131	2024-06-07 15:00:00
Fort William, Kolkata, India	129	2024-06-07 15:00:00
Kolkata US Consulate, India (কোলকাতা আমেরিকী বাণিজ্য দূতাবাস)	124	2024-06-07 15:00:00
Padmapukur, Howrah, India	116	2024-06-07 15:00:00
Ballygunge, Kolkata, India	108	2024-06-07 15:00:00
Belur Math, Howrah, India	106	2024-06-07 15:00:00
Jadavpur, Kolkata, India	105	2024-06-07 15:00:00
Rabindra Bharati University, Kolkata, India	105	2024-06-07 15:00:00
Bidhannagar, Kolkata, India	88	2024-06-07 15:00:00
Rabindra Sarobar, Kolkata, India	77	2024-06-07 15:00:00

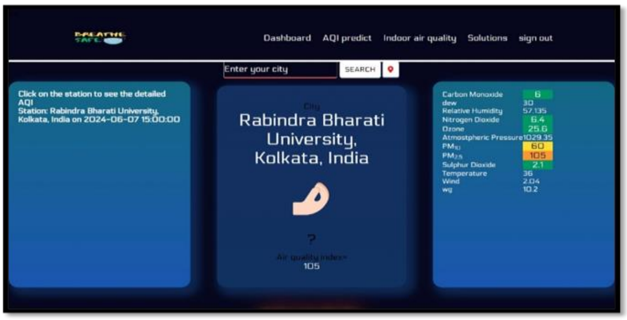
City
Ghusuri, Howrah, India

Terrible
Air quality index-
188

Pollutants & Weather conditions:
Station: Ghusuri, Howrah, India on 2024-06-07 15:00:00

Carbon Monoxide	3.2
dew	30
Relative Humidity	58.22666666666667
Nitrogen Dioxide	8.7
Ozone	14.9
Atmostpheric Pressure	995.0333333333333
PM_{10}	80
$PM_{2.5}$	188
Sulphur Dioxide	2.3
Temperature	36.5
Wind	0.4900000000000005
wg	10.2

Figure 27: Website showing AQI values of a given Location



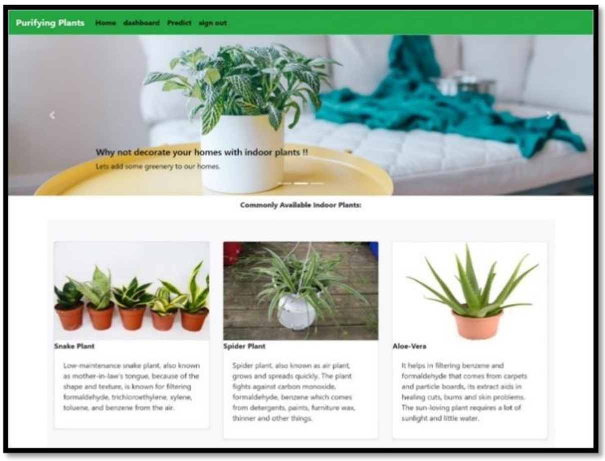
Click on the station to see the detailed AQI
Station: Rabindra Bharati University, Kolkata, India on 2024-06-07 15:00:00

Rabindra Bharati University, Kolkata, India

Air quality index=
105

Carbon Monoxide	6
dew	30
Relative Humidity	57.135
Nitrogen Dioxide	6.4
Ozone	25.6
Atmostpheric Pressure	1029.35
PM_{10}	60
$PM_{2.5}$	105
Sulphur Dioxide	2.1
Temperature	36
Wind	2.04
wg	10.2

Figure 28: Website showing AQI values of Specific Location



Why not decorate your homes with indoor plants !!
Lets add some greenery to our homes.

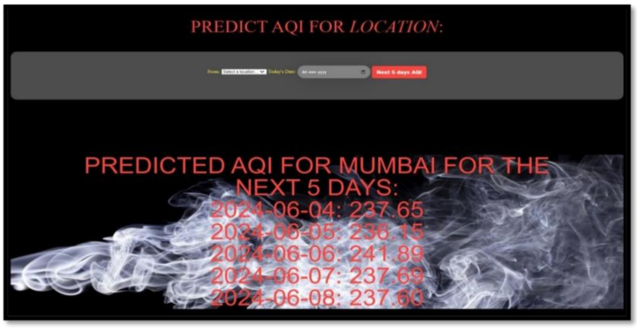
Commonly Available Indoor Plants:

Snake Plant
Low-maintenance snake plant, also known as mother-in-law's tongue, because of the shape and texture, is known for filtering formaldehyde, trichloroethylene, xylene, toluene, and benzene from the air.

Spider Plant
Spider plant, also known as air plant, grows and spreads quickly. The plant fights against carbon monoxide, formaldehyde, benzene which comes from detergents, paints, furniture wax, thinner and other things.

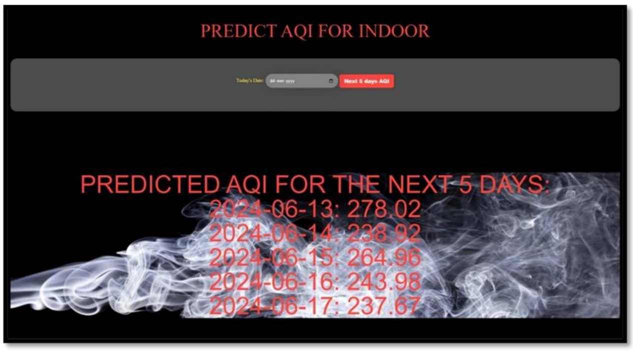
Aloe-Vera
It helps in filtering benzene and formaldehyde that comes from carpets and particle boards, its extract aids in healing cuts, burns and skin problems. The sun-loving plant requires a lot of sunlight and little water.

Figure 29: Website showing precautionary measures to reduce AQI level


PREDICT AQI FOR *LOCATION*:

PREDICTED AQI FOR MUMBAI FOR THE NEXT 5 DAYS:
2024-06-04: 237.65
2024-06-05: 236.15
2024-06-06: 241.89
2024-06-07: 237.69
2024-06-08: 237.60

Figure 30: Website showing Predicted AQI values for the next 5 days from a date for a location (Mumbai)


PREDICT AQI FOR INDOOR

PREDICTED AQI FOR THE NEXT 5 DAYS:
2024-06-13: 278.02
2024-06-14: 238.92
2024-06-15: 264.96
2024-06-16: 243.98
2024-06-17: 237.67

FIGURE 31: Website showing Predicted AQI values for the next 5 days from a date for the place where the device is placed

9. CONCLUSION AND FUTURE WORK

The data which was used in our research is static. However, our proposed model aims to collect hourly data. When the data collection is done in real-time, we can use the cloud which will give us a more accurate performance. The predicted AQI results will be further categorized based on health standards and the air quality measured indoors can be categorized as hazardous or not. Depending on this prediction a website is built where using static data forecasting is done for the next five days. For better usage of this facility app development can be used which will be easier for users to get notifications and alerts. Precautions will be advised depending on the AQI values for example installation of HEPA filters, encouraging usage of N95 masks and planting more indoor plants.

In this study, the performance metrics RMSE and MAPE evaluated for various cities showed that the model PROPHET gave the lowest RMSE and MAPE and thus was observed to be the best fit for forecasting AQI. This model can also be applied to the indoor hardware device setup and thus will help us to achieve accurate forecasting results. If the AQI is predicted correctly pollution can be controlled.

With the help of the prediction and usage of GPS the contaminated area can be determined and also the cause of pollution can be identified. Several pollutants pose a hazardous threat to human health and thus addressing this issue is an urgent matter.

REFERENCES

[1] T. Dutta, A. Dutta, and R. Roy, "Time series analysis of Air Quality Index for several Indian cities," Department of Computer Science, Vidyasagar College, University of Calcutta, Project Report, 2021. [Online]. Available: https://www.academia.edu/104580770/Time_series_analysis_of_Air_Quality_Index_for_several_Indian_cities.

[2] T. C. Bui, V. D. Le, and S. K. Cha, "A deep learning approach for forecasting air pollution in South Korea using LSTM," arXiv preprint arXiv:1804.07891, 2018.

[3] D. Dessai, S. Dessai, and S. D. C. S. e Araujo, "Time Series Analysis for Air Quality Forecasting," Journal of Emerging Technologies and Innovative Research (JETIR), vol. 7, no. 7, pp. 272-277, 2020. [Online]. Available: http://www.jetir.org/papers/JETIR2007331.pdf.

[4] N. K. R, S. Bhumika, S. R, and V. R, "Air Quality Index Prediction using LSTM," International Research Journal of Engineering and Technology (IRJET), vol. 7, no. 6, 2020. [Online]. Available: http://www.irjet.net.

[5] S. Kumar and A. Jasuja, "Air quality monitoring system based on IoT using Raspberry Pi," in 2017 International Conference on Computing, communication and Automation (ICCCA), 2017, pp. 1341-1346.

[6] J. Jo, B. Jo, J. Kim, S. Kim, and W. Han, "Development of an IoT-based indoor air quality monitoring platform," Journal of Sensors, vol. 2020, pp. 1-14, 2020.

[7] A. Barthwal and D. Acharya, "An IoT-based sensing system for modelling and forecasting urban air quality," Wireless Personal Communications, vol. 116, no. 4, pp. 3503-3526, 2021.

[8] R. S. AbdulWahhab, "Air quality system using IoT for indoor environmental monitoring," in Proceedings of the 2019 5th International Conference on Computer and Technology Applications, 2019, pp. 184-188.

[9] K. Rajakumari and V. Priyanka, "Air pollution prediction in smart cities by using machine learning techniques," IJITEE, vol. 9, no. 5, pp. 1272-1279, 2020.

[10] N. Sahu and A. Sarkar, "Time Series Analysis of Air Quality of an Industrial District of India Using Remote Sensing and GIS: Impact of Covid-19 Lockdown," 2024.

[11] N. S. Gupta et al., "Prediction of air quality index using machine learning techniques: a comparative analysis," Journal of Environmental and Public Health, vol. 2023, pp. 1-26, 2023.

[12] M. Imam, S. Adam, S. Dev, and N. Nesa, "Air Quality Monitoring Using Statistical Learning Models for Sustainable Environment," Intelligent Systems with Applications, vol. 200333, 2024.

[13] K. B. Priya Iyer and V. Dharshini, "ANALYSIS AND PREDICTION OF AIR QUALITY INDEX IN INDIA DURING PRE AND POST COVID PANDEMIC USING MACHINE LEARNING ALGORITHMS," PalArch's Journal of Archaeology of Egypt / Egyptology, vol. 17, no. 7, pp. 6995-7003, 2020. [Online]. Available: https://archives.palarch.nl/index.php/jae/article/view/3097.

[14] R. M. Patil, H. T. Dinde, S. K. Powar, and P. M. Ganeshkhind, "A literature review on prediction of air quality index and forecasting ambient air pollutants using machine learning algorithms," Int J Innov Sci Res Technol, vol. 5, no. 8, pp. 1148-1152, 2020.

[15] H. Patel and S. Saket, "Air pollution prediction system for the smart city using data mining technique: a survey," Health, vol. 6, no. 12, 2019.

[16] R. Sharma, G. Shilimkar, and S. Pisal, "Air quality prediction by machine learning," Int. J. Sci. Res. Sci. Technol, vol. 8, pp. 486-492, 2021.

[17] A. Choudhary et al., "Time Series Simulation and Forecasting of Air Quality Using In-situ and Satellite-Based Observations Over an Urban Region," Nature Environment & Pollution Technology, vol. 21, no. 3, 2022. [Online]. Available: https://doi.org/10.46488/NEPT.2022.v21i03.018.

[18] V. Chaudhary, A. Deshbhratar, V. Kumar, and D. Paul, "Time series-based LSTM model to predict air pollutant's concentration for prominent cities in India," UDM, Aug. 2018. [Online]. Available: https://philippe-fournier-viger.com.

[19] V. Deva Sekhar and P. Natarajan, "Prediction of air quality and pollution using statistical methods and machine learning techniques," International Journal of Advanced Computer Science and Applications, vol. 14, no. 4, 2023. [Online]. Available: https://proquest.com.

[20] J. Kalajdjieski, M. Korunoski, B. R. Stojkoska, and K. Trivodaliev, "Smart city air pollution monitoring and prediction: A case study of Skopje," in ICT Innovations 2020. Machine Learning and Applications: 12th International Conference, ICT Innovations 2020, Skopje, North Macedonia, September 24–26, 2020, Proceedings 12. Springer International Publishing, 2020, pp. 15-27. [Online]. Available: https://researchgate.net.