Dissertation on

Monitoring the concentration of air pollutants and its health hazards using Machine Learning models

Submitted in partial fulfilment of the requirements for the award of degree of

Bachelor of Technology
in
Computer Science & Engineering

UE20CS390A – Capstone Project Phase - 1

Submitted by:

Aditi Jain	**PES2UG20CS021**
Aditya R Shenoy	**PES2UG20CS025**
Ananya Adiga	**PES2UG20CS043**
Anirudha Anekal	**PES2UG20CS051**

Under the guidance of

Prof. Saritha R
Assistant Professor
PES University

January - May 2023

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

FACULTY OF ENGINEERING

CERTIFICATE

This is to certify that the dissertation entitled

Monitoring the concentration of air pollutants and its health hazards using Machine Learning models

is a bonafide work carried out by

Aditi Jain	**PES2UG20CS021**
Aditya R Shenoy	**PES2UG20CS025**
Ananya Adiga	**PES2UG20CS043**
Anirudha Anekal	**PES2UG20CS051**

In partial fulfilment for the completion of sixth semester Capstone Project Phase - 1 (UE20CS390A) in the Program of Study -Bachelor of Technology in Computer Science and Engineering under rules and regulations of PES University, Bengaluru during the period Jan. 2023 – May. 2023. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 6th semester academic requirements in respect of project work.

Signature	Signature	Signature
Prof. Saritha R	Dr. Sandesh B J	Dr. B K Keshavan
Assistant Professor	Chairperson	Dean of Faculty

External Viva

Name of the Examiners **Signature with Date**

1. Dr. Sarasvathi V ____________________________

2. Prof.Komal Baheti ____________________________

DECLARATION

We hereby declare that the Capstone Project Phase - 1 entitled **Monitoring the concentration of air pollutants and its health hazards using Machine Learning models** has been carried out by us under the guidance of Prof. Saritha R, Assistant Professor, PES University and submitted in partial fulfilment of the course requirements for the award of degree of **Bachelor of Technology** in **Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester January – May 2023. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

Name	**SRN**	**Signature**
Aditi Jain	PES2UG20CS021	
Aditya R Shenoy	PES2UG20CS025	
Ananya Adiga	PES2UG20CS043	
Anirudha Anekal	PES2UG20CS051	

ACKNOWLEDGEMENT

I would like to express my gratitude to Prof. Saritha R, Department of Computer Science and Engineering, PES University, for her continuous guidance, assistance, and encouragement throughout the development of this UE20CS390A - Capstone Project Phase – 1.

I am grateful to the Capstone Project Coordinators, Dr.Sarasvathi V, Professor and Dr. Sudeepa Roy Dey, Associate Professor, for organizing, managing, and helping with the entire process.

I take this opportunity to thank Dr. Sandesh B J, Professor & Chairperson, Department of Computer Science and Engineering, PES University, for all the knowledge and support I have received from the department. I would like to thank Dr. B.K. Keshavan, Dean of Faculty, PES University for his help.

I am deeply grateful to Dr. M. R. Doreswamy, Chancellor, PES University, Prof. Jawahar Doreswamy, Pro Chancellor – PES University, Dr. Suryaprasad J, Vice-Chancellor, PES University for providing to me various opportunities and enlightenment every step of the way. Finally, this Capstone Project could not have been completed without the continual support and encouragement I have received from my family and friends.

ABSTRACT

Industrialization is a leading cause of pollution in today's world, and a lack of awareness of air quality has led to people being unconscious of a threat that looms over them with every breath they take. PM2.5, PM10, NO2, SO2, and CO2 are examples of pollutants which are extremely harmful to human health, especially the lungs, since they are the first to come into contact when we breathe in air. It is vital for the masses to be mindful of these health hazards.

Thus, we propose this system to continuously monitor the air quality around the user and alert them of the probability of their contracting lung cancer. This project is primarily a research based one where we aim to improve the efficiency of already existing systems so that they can be used in the current time with scalability to the current amount of data for maximum efficiency.

Our system aims to use a hybrid model of Adaptive LSTM and ARIMA models. These both have been proven to work well with time series data, and they are also easy to implement. They have high reliability and can dynamically adjust to changes.

After the model has been trained, we will then deploy it on a safe and secure cloud platform to ensure scalability, making it easy for users to access it and making it relevant to present times.

We will make use of Arduino boards and high quality sensors to collect real time data about the air quality in a particular environment. This data will be used for predictions and fed back to the model to learn from. The model will thus train constantly on real time and static data, helping it make better predictions over time.

Our main objective is to spread awareness and increase accuracy in predictions that can potentially lead to saving people's lives and the government and industry making environmentally conscious decisions for future generations.

Table of Contents

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Motivation

The industrialization process has been a significant contributor to the current pollution crisis facing the world. However, many people remain unaware of the dangers associated with poor air quality. PM2.5, PM10, NO2, SO2, and CO2 are examples of pollutants which are particularly harmful to human health, with the lungs being the first point of contact. Particulate matter of a small diameter enters the lungs easily. As such, it is crucial to increase public awareness of these hazards and encourage individuals to take protective measures.

To address this issue, it is necessary to develop and implement air quality monitoring systems. Such systems would provide individuals with real-time information on air quality levels in their surroundings and alert them to potential health risks. By increasing awareness and providing early warnings, we can help individuals take preventative measures and protect their health.

1.2 Solution

Therefore, we propose a continuous air quality monitoring system that will keep track of the air quality in the user's vicinity and alert them to the likelihood of developing lung cancer. Our research-based project aims to improve the effectiveness of existing systems to ensure maximum efficiency, taking into account the current volume of data.

The proposed system is designed to monitor the levels of PM2.5, PM10, NO2, and CO in the air continuously, providing the user with real-time information about the quality of the air they are breathing. The system will then use this data to assess the user's risk of developing lung cancer and alert them accordingly. By keeping track of the air quality and providing early warnings, this system can help people take appropriate measures to safeguard their health.

In conclusion, the proposed air quality monitoring system is an essential tool for ensuring people's health and well-being in today's polluted environment. With its ability to continuously monitor the air quality and provide real-time alerts, this system will enable people to take preventive measures to protect their health from the harmful effects of air pollution.

CHAPTER II

PROBLEM DEFINITION

2.1 Problem Statement

To address the issue of increasing mortality rate due to air pollution , we propose a hybrid model of Adaptive LSTM and ARIMA deep learning models that can predict the probability of a person being affected by the air quality in his surroundings.

The model will be deployed on a cloud platform which ensures scalability and accessibility to store huge amounts of real time data of air quality collected using arduino boards and high quality IOT sensors. Continuously monitoring air quality, predicting and alerting individuals to potential health risks. These deep learning models have proven to be effective in handling time series data and are easy to implement. Moreover, they have high reliability and can dynamically adjust to changes, making them ideal for real-time monitoring of air quality.

By constantly training on real-time and static data, the model will become increasingly accurate, allowing it to provide more reliable and relevant predictions. This will enable individuals to take preventive measures to protect themselves from the harmful effects of air pollution and make informed decisions about their daily activities.

CHAPTER III

LITERATURE SURVEY

3.1 Determination of Air Quality Life Index (AQLI) in Medinipur City of West Bengal(India) During 2019 To 2020 : A contextual Study

3.1.1 Approach & Results:

- They do not use any ML model. They use the predefined method provided by various organizations like WHO, INAAQS etc.
- They show a provable reduction in life expectancy and the correlation to various air pollutants.
- They also speak about the direct effects of PM2.5 on the lungs.

3.1.2 Advantages:

- They use an air pollutant rather than AQI.
- Speaks about the direct effects of PM on various parts of the body.
- Uses an expensive but very accurate, and lab tested/ approved sensor for PM 2.5
- (Plantower particulate matter sensor PMS3003)

3.1.3 Limitations:

- No ML model used. They use existing methods to determine everything.

- Only looks at one pollutant.
- Not clear if all the effects (of the pollutant) are considered .

3.2 The nexus between COVID-19 deaths, air pollution and economic growth in New York state: Evidence from Deep Machine Learning.

3.2.1 Approach & Results:

- They use an ensemble of ML models (ANN, Deep Learning, Decision Tree)
- The association between COVID-19-related mortality, economic growth, PM10, PM2.5, and NO2 concentrations in New York state is examined in this article.

3.2.2 Advantages:

- They consider individual pollutants.
- Very well thought out data gathering and cleaning.
- The use methods used are well suited for the type of predictions and the type of data we are using.

3.2.3 Limitations:

- It has been done mainly to covid related deaths.
- This has many inherent issues especially considering how little we truly know about covid.
- They attribute over 50% of deaths due to covid as deaths facilitated by these pollutants.
- This cannot be proven due to the lack of research time on the effects of covid.

- Does not consider other underlying issues

3.3 An Application of IoT and Machine Learning to Air Pollution Monitoring in Smart Cities

3.3.1 Approach & Results:

- They use ANN models for predicting the concentrations of air contaminants and Pearson's coefficient for the correlation with weather conditions.
- The objective of this paper is to monitor pollutant concentrations in smart cities and find a correlation between pollutants and weather parameters and prediction of the pollutants to prevent diseases like lung cancer.

3.3.2 Advantages:

- The paper takes into account the weather conditions as well. As air pollutants mixed with other factors like water/high winds cause differing effects
- The model created has achieved a Root Mean Square Error of only 0.0128 for SO2 prediction and 0.0001 for PM2.5
- They have used two different methods (Pearson Correlation and ANN) for the weather correlation and prediction.

3.3.3 Limitations:

- Does not consider all the environmental weather conditions (Only wind speeds, temperature and humidity)

- The model only contains a single hidden layer and they have not run it only for 500 epochs
- The IoT infrastructure they have used is very crowded and complex.

3.4 Mohammad Marzouk and Mohammad Atef "Assessment of indoor air quality in academic buildings using IOT and deep learnings": Mdpi(June 2022)

3.4.1 Approach & Results:

To find the correlation between outdoor pollution and indoor air quality, by monitoring real time IAQ using IOT sensors and sending the collected data to cloud via wireless connections.CNN and LSTM are used to train the collected data.

3.4.2 Advantages:

- Uses real time data.
- Finds the correlation between outdoor and indoor environments.
- Concentrates on rural areas too, which might be affected by burning wood, construction particles and unpaved roads.

3.4.3 Limitations:

- Does not consider the family history with lung cancers.
- Does not consider the occupational exposures to carcinogens.

- It is biased as the readings were collected only during summer when the humidity will be relatively high.
- The study was restricted to just a few primary pollutants and did not consider volatile organic compounds, NH3 and O3

3.5 Jin Z-Y, Wu M, Han R-Q, Zhang X-F, Wang X-S, et al. (2014) "Household Ventilation May Reduce Effects of Indoor Air Pollutants for Prevention of Lung Cancer: A Case-Control Study in a Chinese Population. PLoS ONE 9(7): e102685. doi:10.1371/journal.pone.0 102685

3.5.1 Approach & Results:

The purpose of this research is to investigate the link between home ventilation and lung cancer.
A standardized questionnaire was used to obtain epidemiologic and home ventilation data.
The adjusted odds ratios (ORadj) and 95% confidence intervals (CI) were calculated using unconstrained logistic regression.

3.5.2 Advantages:

- The information gathered takes into account a person's family history of lung cancer, basic demographic characteristics, socioeconomic status, tobacco smoking history, alcohol intake, dietary history, and physical activity.

- Active smokers, second hand smokers, carcinogens, tobacco smoking and high temperature cooking oil flames were considered for the study thereby covering a wide scope.

3.5.3 Limitations:

- Because the data is limited to the local community, selection and memory bias may present in the study.
- The data about family history, age, gender is considered via a standardized questionnaire by interviewers, using quantitative data might have been effective instead.
- Have not considered occupational exposure
- The study is limited to IAQ and does not consider how outdoor pollutants are affecting IAQ.

3.6 How Is the Lung Cancer Incidence Rate Associated with Environmental Risks? Machine-Learning-Based Modeling and Benchmarking

3.6.1 Approach & Results:

- They use an ensemble of models like Logistic Regression, Random Forest, SVM and Gradient Boosting and perform benchmarking on them.
- The goal of this research is to use machine-learning-based modeling and benchmarking to examine the association between lung cancer incidence rate and air pollution.
- The study's goal is to create a prediction model to better understand the influence of air pollution on illness.

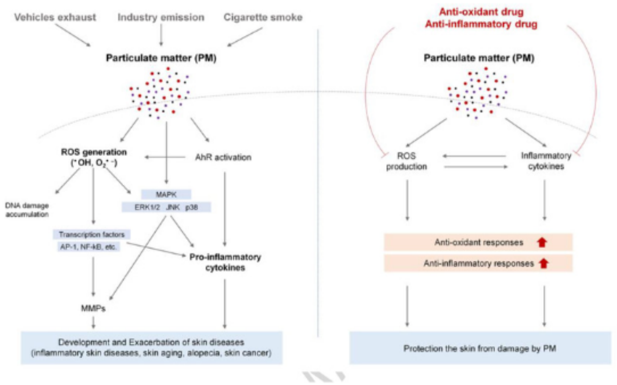
3.6.2 Advantages:

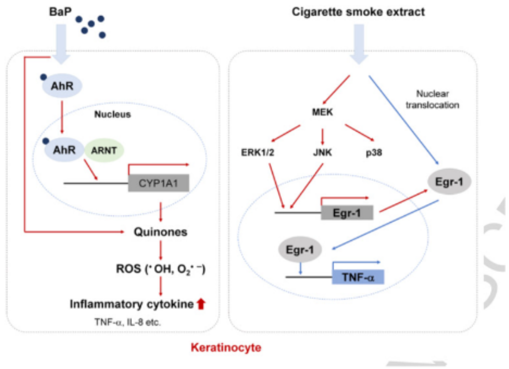
- They perform benchmarking, i.e, comparing the performance of different ML models and provides insights on which the best one is .
- The dataset they have used contains both environmental risk factors (air pollutants) and the lung cancer incidence rates from various countries and hence increases generalizability.

3.6.3 Limitations:

- Does not consider all the environmental weather conditions (Only wind speeds, temperature and humidity)
- The model only contains a single hidden layer and they have not run it only for 500 epochs
- The IoT infrastructure they have used is very crowded and complex.

3.7 Kim Kyung Eun, Cho Daeho, Park Hyun Jeong, Air pollution and skin diseases: Adverse effects of airborne particulate matter on various skin diseases, Life Sciences (2016), doi: 10.1016/j.lfs.2016.03.039





3.7.1 Approach & Results:

- This page discusses the relationship between PM and skin illnesses, as well as the underlying immunological pathways.
- Increased PM levels have been linked to the development of a variety of skin illnesses through the control of oxidative stress and inflammatory cytokines.
- As a result, antioxidant and anti-inflammatory medicines may be beneficial in the treatment of PMS-induced skin disorders.

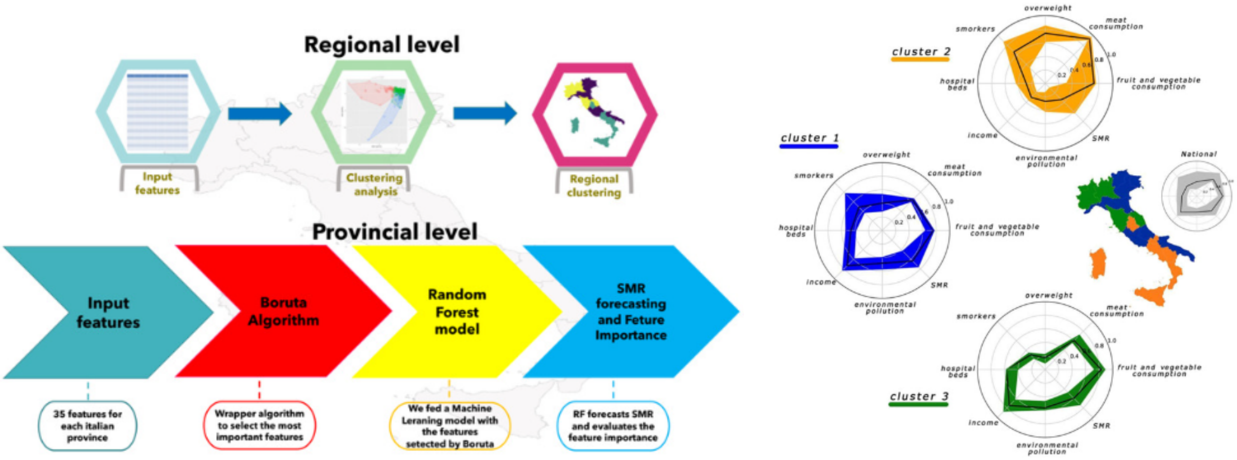
3.7.2 Advantages:

- This paper has listed in detail which air pollutants are known to cause which disease.
- It lists various diseases like Atopic dermatitis, Acne, Psoriasis, and skin cancer.
- It also discusses the conditions of skin aging, alopecia, and oxidative stress.

3.7.3 Limitations:

- The amount of statistics is less than is used to prove its point.
- It doesn't discuss the correlation between these diseases, and how they affect the probability of another happening.

3.8 Roberto Cazzolla Gatti, Arianna Di Paola, Alfonso Monaco, Alena Velichevskaya, Nicola Amoroso, Roberto Bellotti, The spatial association between environmental pollution and long-term cancer mortality in Italy, Science of The Total Environment, https://doi.org/10.1016/j.scitotenv. 2022.158439


Regional level
Input features
Clustering analysis
Regional clustering
Provincial level
Input features
Boruta Algorithm
Random Forest model
SMR forecasting and Feture Importance
35 features for each italian province
Wrapper algorithm to select the most important features
We fed a Machine Leraning model with the features setected by Boruta
RF forecasts SMR and evaluates the feature importance
cluster 2
cluster 1
cluster 3
National
overweight
meat consumption
fruit and vegetable consumption
SMR
environmental pollution
income
hospital beds
smorkers
0.2
0.4
0.6
0.8
1.0

3.8.1 Approach & Results:

- Using an artificial intelligence method, this article examined the relationships between cancer mortality, socioeconomic characteristics, and sources of environmental pollution in Italy at both broad regional and narrower provincial scales.
- Random Forest (RF) regression in combination with a Boruta feature importance analysis, K means clustering.
- SMR forecasting and Feature importance, Regional cluster analysis.

3.8.2 Advantages:

- It has taken into consideration all the different types of body parts that can get affected due to air pollution in extensive detail.
- The data sources and the algorithms used have been discussed.
- Investigated the possible geographical relationship between socioeconomic and lifestyle characteristics.

3.8.3 Limitations:

- Some individual sources of pollution have no link to any form of cancer. This study is limited to Italy, but it should be expanded to other nations as well, because air pollution levels vary by region.

3.9 Conclusion to the literature survey:

Most studies only include a small number of risk variables when examining the influence of air quality on human health. Some initiatives, for example, consider just indoor air quality (IAQ) or outdoor air quality (OAQ), but not both, or they exclude environmental variables, family history, and occupational exposure. This method can lead to biased and untrustworthy results that do not completely reflect the complexities of the topic at hand.

Furthermore, despite the fact that air pollution is a major public health problem, little study has been conducted on its real health impacts. While numerous research have looked at the association between air quality and health outcomes, few have looked at the causative relationship. Due to a dearth of study, properly predicting the influence of air quality on health can be difficult.

Another problem is that projections based on static and out-of-date data are uncommon. This can be an issue since air quality can fluctuate fast owing to a number of variables such as weather, industrial activity, and transportation. Predictions based on old data might result in incorrect conclusions and perhaps dangerous choices.

To address these difficulties, more extensive and sophisticated models that reflect a larger variety of risk variables and the real health impacts of air pollution are required. These models must also be based on real-time data and capable of fast adapting to changes in air quality levels. This allows us to create more precise forecasts regarding the impact of air pollution on human health, which may guide policy decisions and assist individuals in making educated decisions about their health and well-being.

CHAPTER IV

DATA

4.1 Overview

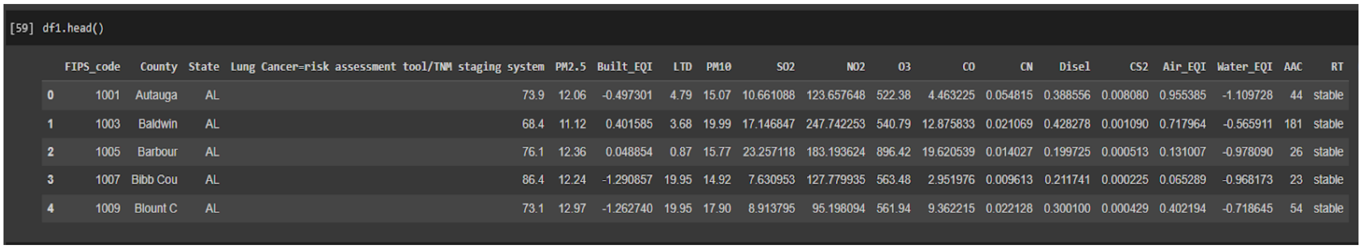
We aim to use a static dataset containing correlation between lung cancer and air pollutants, for training the model.The dataset has been obtained from a reliable website "Harvard dataverse", by Harvard University.

The dataset is available at:

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HMOEJO

We will be collecting real time air quality data using IOT sensors for prediction.

4.2 Dataset



```
[59] df1.head()
```

	FIPS_code	County	State	Lung Cancer=risk assessment tool/TNM staging system	PM2.5	Built_EQI	LTD	PM10	SO2	NO2	O3	CO	CN	Disel	CS2	Air_EQI	Water_EQI	AAC	RT
0	1001	Autauga	AL	73.9	12.06	-0.497301	4.79	15.07	10.661088	123.657648	522.38	4.463225	0.054815	0.388556	0.008080	0.955385	-1.109728	44	stable
1	1003	Baldwin	AL	68.4	11.12	0.401585	3.68	19.99	17.146847	247.742253	540.79	12.875833	0.021069	0.428278	0.001090	0.717964	-0.565911	181	stable
2	1005	Barbour	AL	76.1	12.36	0.048854	0.87	15.77	23.257118	183.193624	896.42	19.620539	0.014027	0.199725	0.000513	0.131007	-0.978090	26	stable
3	1007	Bibb Cou	AL	86.4	12.24	-1.290857	19.95	14.92	7.630953	127.779935	563.48	2.951976	0.009613	0.211741	0.000225	0.065289	-0.968173	23	stable
4	1009	Blount C	AL	73.1	12.97	-1.262740	19.95	17.90	8.913795	95.198094	561.94	9.362215	0.022128	0.300100	0.000429	0.402194	-0.718645	54	stable

The dataset contains 2603 rows and 19 attributes. The attributes are-

Attributes	Attribute description
FIPS_code	A code that uniquely identifies the geographic area
County	The county in US, where the air quality was monitored
State	The state in that county where the air quality was monitored
Lung cancer	Measured in terms of risk assessment tool
Pollutants- PM2.5, NO2, PM10,SO2,O3,CO,CN,CS2	Rate of pollutants present in air
Built_EQI	The quality of residential environment
Air_EQI	Overall air quality
Water EQI	Overall quality of water in that area
AAC	The age adjusted incidence rate refers to the rate of new cases of lung cancer diagnosed in a population after accounting for the population's age distribution..
RT- radiation therapy	Both non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) are commonly treated with radiation therapy.
LTD- - limited disease	LTD- it refers to a stage of the disease in which cancer is confined to one lung and possibly nearby lymph nodes.

.

CHAPTER V

SYSTEM REQUIREMENTS SPECIFICATION

5.1 Introduction

In response to the growing concern over the dangers of air pollution on human health, we propose a hybrid model of Adaptive LSTM and ARIMA models to monitor air quality continuously. These models have proven to be highly reliable and adaptable to changes in time series data. To ensure scalability, we plan to deploy the model on a secure cloud platform, making it easily accessible to users. Real-time data will be collected using Arduino boards and high-quality sensors, allowing the model to learn continuously and improve its predictions over time. By increasing accuracy in predictions and spreading awareness, we aim to encourage environmentally conscious decision-making and potentially save lives.

5.2 System Requirements

5.2.1 Features

The product features entail:

- The user will be able to check the current levels of pollution in his surroundings and the probability of contracting Lung cancer and Skin cancer based on the current exposure to pollutants.

- The user will receive a notification if the concentration of the pollutants crosses a threshold value and becomes hazardous.

5.2.2 User classes and characteristics

- **Smart home companies:**

  Smart home companies currently incorporate various sensors like temperature sensors, heat sensors and motion sensors. They can also incorporate our pollutant sensors as an additional layer of detection on the current infrastructure.

- **Government:**

  Governments can use our product to identify pollution hotspots and create policies and regulations to reduce the harmful emissions and issue public health warnings in high risk areas.

- **City planning:**

  City planners and private companies can use this information to guide urban planning and development and create communities in a healthier living space.

- **General Public:**

  The general should be informed about the levels of pollution and the risks associated with it to take preventive measures

5.2.3 Operating Environment

- **Arduino:**

  The default operating system present on the board will be used. We will add our custom code to it. Any OS that supports the Arduino IDE (or the web version) can be used to interface with the Arduino (this is only to add code to the board).

- **Coding Platform:**

  Any OS and any editor can be used that you prefer as long as it supports git or some other way to have a shared codebase. It should also have inbuilt support for languages that will be used for coding and a way to integrate databases.

5.2.4 Software requirements

- **Data collection and preprocessing tools**:

  We will need tools to collect air pollutant data from various sources, such as air quality sensors. We will also need software to clean, normalize, and prepare the data for analysis.

- **Machine Learning frameworks:**

  We will use machine learning frameworks such as TensorFlow, Keras, or PyTorch to build and train our models. These frameworks provide a range of tools for data processing, model building, and evaluation.

- **Statistical analysis tools:**

  We will need statistical analysis tools such as R or Python packages like NumPy and Pandas for data analysis and exploratory data analysis (EDA).

- **Visualization tools:**

  We will need visualization tools like Matplotlib, Seaborn, or Tableau to create visualizations and graphs that help you understand the data and communicate results.

- **Cloud computing platforms:** To store and analyse massive amounts of data and deploy our machine learning models, we will utilize a cloud computing platform such as AWS, GCP, ThingSpeak, or Azure.

- **Database management systems:** To store and manage the acquired data, we'll need a database management system like MySQL, PostgreSQL, or MongoDB.

5.2.5 Hardware requirements

- ESP8286, a wifi module, will be used to store the collected real time data.
- MQ135 to monitor NO2
- Dust sensor to monitor particulate matter of different aerodynamic diameter

5.2.6 User Interface

- Basic information can be displayed on an LCD connected to the stations. More data can be viewed on a computer with basic internet connection (either through a custom made website/ application or the actual cloud platform directly)
- In our provided interface there will be basic instructions on how to interpret the data received.
- The stations can measure every . It can send each measurement to the database, preferably immediately. The model can rerun with the new received data every .
- All error messages will be sent to the cloud platform. Some errors can be displayed on the LCD.

5.2.7 Communication interface

- **Arduino:**

  The default operating system present on the board will be used. We will add our custom code to it. Any OS that supports the arduino IDE (or the web version) can be used to interface with the arduino (this is only to add code to the board) .
- **Cloud:**

  We will use a few cloud platforms for this. This will evolve as time progresses and we find the pros and cons of each platform. Thinkspeak to save the data from the sensor stations, AWS Glue, Google Cloud Dataflow or Azure Data Factory to process our data and to host our model.

5.3 Functional Requirements

- **Machine Learning Algorithms:** Implement ML algorithms to analyze the air quality data and find the correlations between pollutants and lung cancer rates.
- **Cloud Infrastructure:** Use a cloud infrastructure such as AWS, Google Cloud Platform, or Microsoft Azure to host the ML models and the database.
- **Deployment:** Deploy the ML models to the cloud infrastructure and provide a userfriendly interface to allow users to access the models and the data.
- **Real-time Monitoring:** Continuously monitor the air quality using IoT devices and update the database in real-time. The ML models should also be updated periodically to ensure that they remain accurate.
- **Alert System:** The system should send alerts if the concentration of pollutants exceeds a certain threshold.
- **Disease Prediction:** Based on the quantity of contaminants in the air, the system should be able to assess the collected data and compute the likelihood of lung cancer and other associated diseases.

5.4 Non-Functional Requirements

- **Performance:** The system should be able to manage enormous volumes of data and process it in real time.
- **Maintainability:** The system should be simple to maintain and upgrade, with as little downtime as possible.
- **Security:** Because IoT devices are frequently attacked, the system must be secure and preserve data privacy.
- **Scalability:** The system should be scalable and capable of supporting an increasing number of devices and data points.

5.5 Data Requirements

- **Data volume:** Sufficient data volume is required to train the machine learning model effectively. Generally, more data leads to better accuracy and more robust models.
- **Data quality:** The data used in machine learning projects must be accurate and representative of the real-world scenarios that the model will encounter. Inaccurate or inconsistent data can lead to incorrect predictions and unreliable models.

- **Data diversity:** It is essential for ensuring that the model is not biased toward a certain sample of data. To generalize properly, the model needs to be exposed to a variety of data samples.
- **Data labeling:** Labeled data is required to train supervised learning models effectively. The labeling process involves adding tags or labels to the data that indicate the input/output relationship.
- **Data preprocessing:** Preprocessing involves cleaning and transforming the raw data to prepare it for machine learning algorithms. Data preprocessing can include tasks such as normalization, feature engineering, and data augmentation.
- **Data storage and management:** to handle massive amounts of data and make it available to machine learning algorithms.

CHAPTER VI

SYSTEM DESIGN

6.1 Current System

Current systems have many shortcomings that are listed as follows:

- They monitor air quality but do not predict the health effects of the monitored air on the human body.

  This is extremely harmful as the ignorant user will not be able to generate any useful inferences from it making the whole system redundant.

- They have modeled LSTM or just random forest algorithms. This does not provide proper accuracy or efficiency.

  These models are not scalable for real time data or upcoming environmental factors.

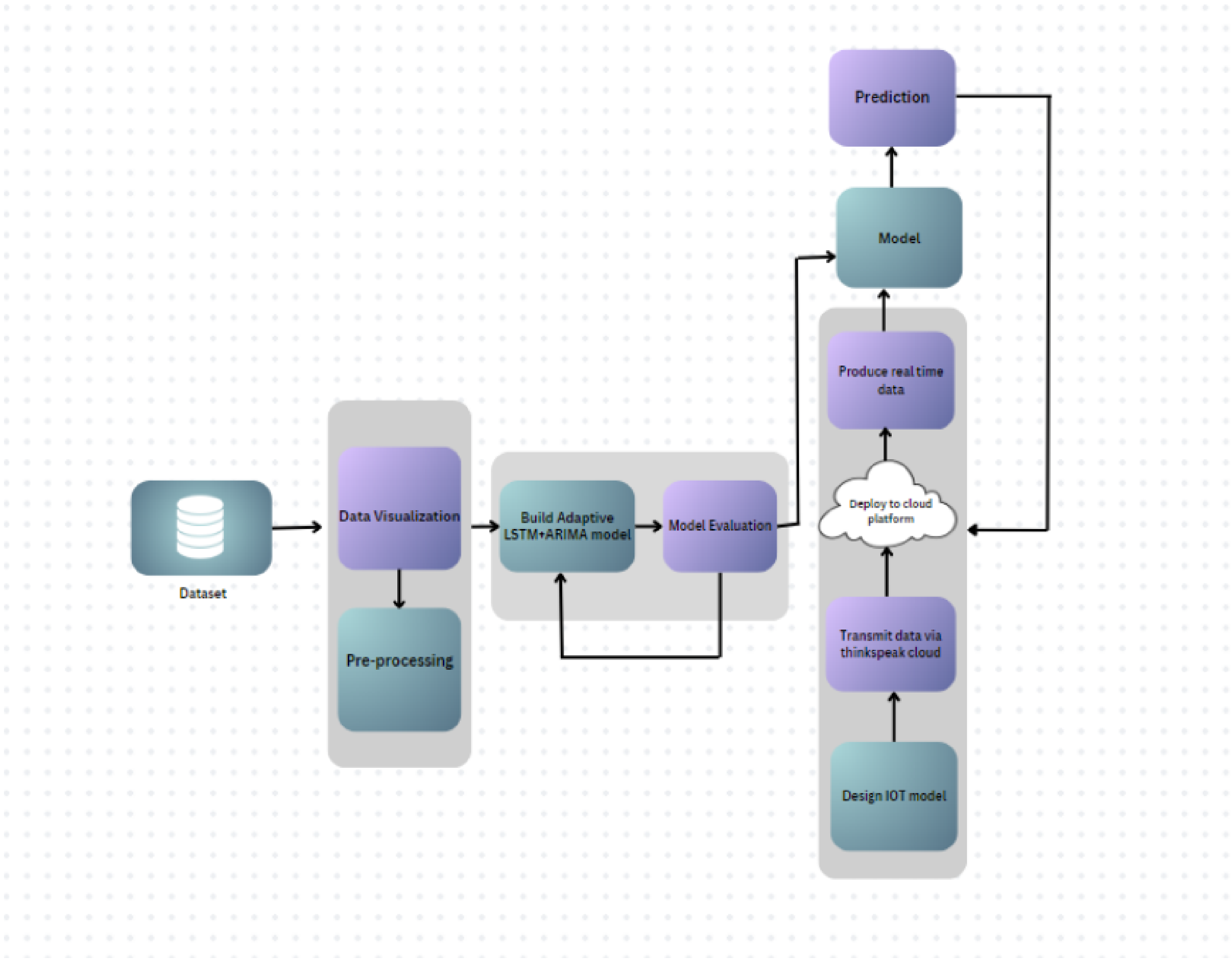
- They focus on any one factor that affects lung cancer.

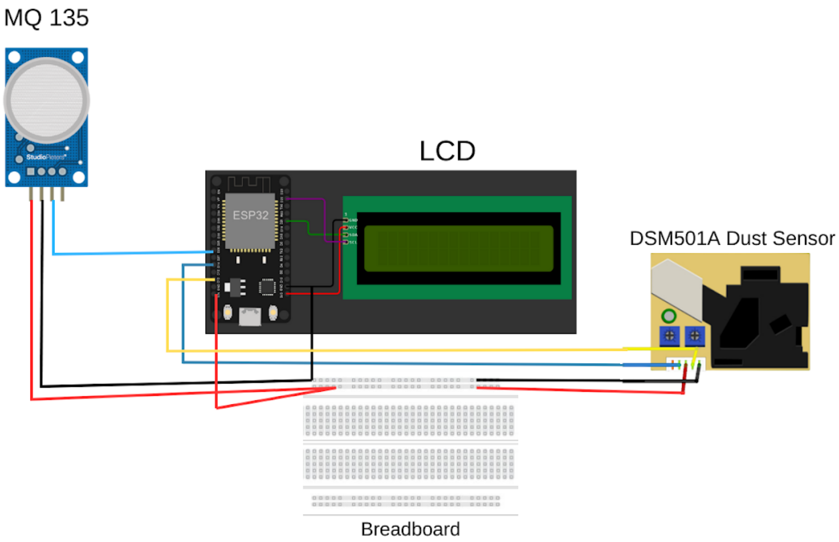
  They do not consider that there are multiple air pollutants that lead to lung cancer or lung cancer related diseases. Our research has shown that PM2.5, PM10, NO2, SO2, and CO all cause lung cancer.

- Furthermore, they have not used any cloud based technologies to test the project's feasibility. We will use these platforms to make the system more scalable and reliable.

6.2 Design Considerations

6.2.1 Architecture:





We will be using 2 models of ESP boards. ESP 32 and ESP 8266. The wiring diagrams for both are very similar and need no extra effort to switch from one board to the other. The dust sensor used can be changed based on availability and price. We are considering DSM501A Dust Sensor, GP2Y1010AU0F and PMS5003. For the sake of the circuit diagram we will use DSM501A.

We have 3 components wired up to the board.

- **LCD:** SDA(data pin) is connected to D21 and the SCL(clock) is connected to D22. It will receive power directly from ESP.
- **MQ135**: It receives power from the power line from ESP to the breadboard. Ground is set up in the same way. The digital output from it is connected to D26.
- **DSM501A**: It receives power from the power line from ESP to the breadboard. Ground is set up in the same way. It has 2 outputs. PM 1 output to D14 and PM 2.5 output to D13.

6.2.2 Detailed flow

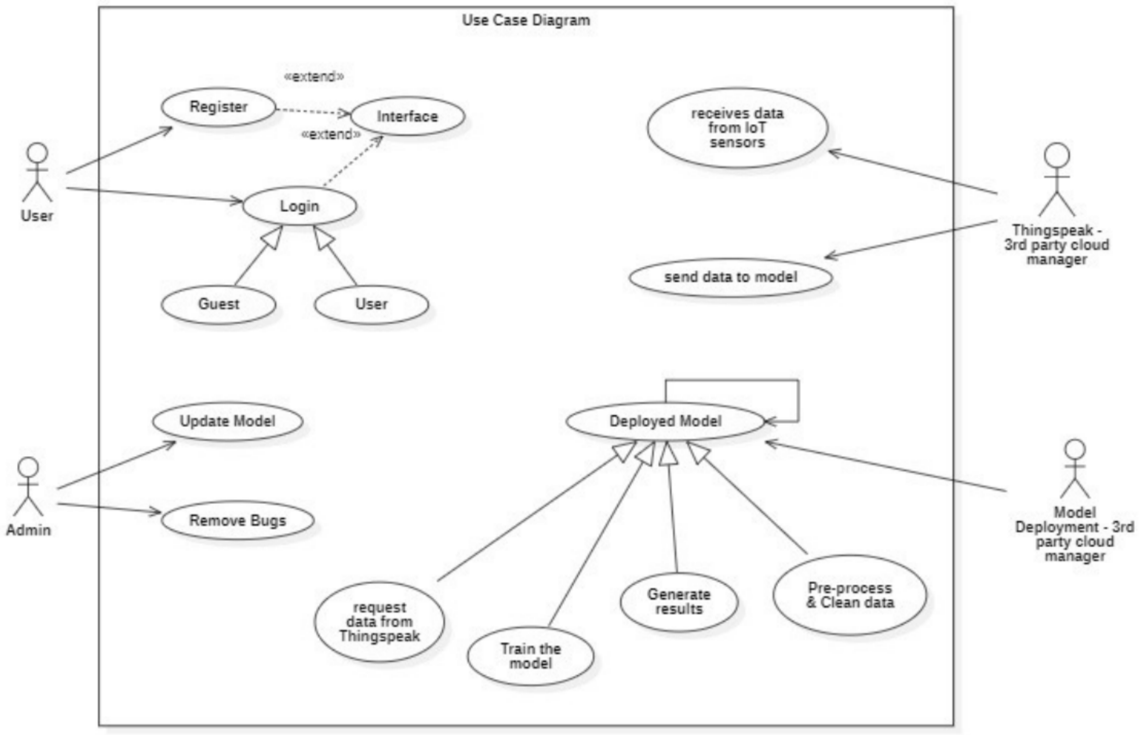
- **Pre-process the data**: Clean and preprocess your time series data, including handling missing values, transforming variables if necessary, and splitting the data into training and testing sets.
- **Train ARIMA model:** Fit an ARIMA model to the training data to capture the seasonal and trend components of the time series. Tune the ARIMA hyperparameters (e.g., order of differencing, autoregressive and moving average orders) using techniques like grid search or time series cross-validation.
- **Generate ARIMA forecasts**: Use the trained ARIMA model to generate short-term forecasts on the testing data.
- **Compute ARIMA residuals**: Calculate the residuals, which are the differences between the actual values and the ARIMA forecasts, for the training and testing data.
- **Train Adaptive LSTM model**: Use the ARIMA residuals as input to train an Adaptive LSTM model on the training data, capturing the residual patterns that ARIMA may have missed.
- **Combine forecasts**: Generate forecasts using the trained Adaptive LSTM model on the testing data and combine them with the ARIMA forecasts to obtain the final hybrid forecast.
- **Evaluate and optimize**: Evaluate the hybrid ARIMA-Adaptive LSTM model's performance on the testing data using appropriate evaluation metrics, and fine-tune the model as needed by altering hyperparameters or model architecture.
- **Use sensors to get real time data:** We will use ESP32 / ESP 8266 modules to construct the IOT sensor stations. Both of them are capable of connecting to the internet. MQ 135 will be used to detect CO, CO2, and NH4. GP2Y1010AU0F will be used for dust detection.
- **Deploy the model and sensor station to cloud platforms:** such as Thingspeak and Streamlit, Spaces or AWS to host the machine learning model.

6.2.3 Pros and Cons of Using the Hybrid Network:

PROS	**CONS**
Both ARIMA and adaptive LSTM work best with time series data.	Adaptive LSTM limits the accuracy in cases where there is limited data.
ARIMA is easy to interpret and reliable for small and stationary datasets.	Adaptive LSTM models can be highly complex, making them difficult to interpret.
While adaptive LSTM works best for large, real time, varying length, non-stationary data and learns complex patterns.	Selecting the appropriate ARIMA model parameters (such as the order of differencing and the number of autoregressive and moving average terms) can be challenging and require a good understanding of the underlying data.
Adaptive LSTM is capable of dynamically adjusting to changes.	
LSTMs can capture long-term dependencies and temporal patterns in the data, which can be useful for predicting lung cancer risk based on changing air quality conditions.	
ARIMA captures the short-term forecasts, and adaptive LSTM captures the long-term forecasts.	

6.3 DESIGN DESCRIPTION:

6.3.1 Use case diagram:



Use cases we have identified throughout our research:

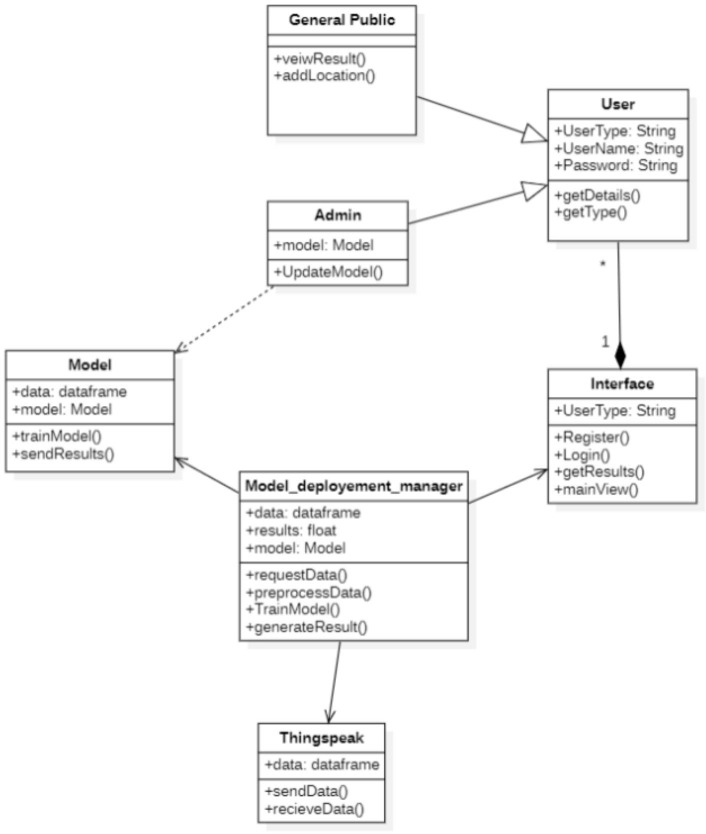
- **User:** User will be an individual or entity that interacts with a system to see the predictions done on the air quality around them. User will register and/or login in the system and will be the source of data from which the model will be further trained.
- **Admin:** Admin will be responsible for managing and maintaining a system or application. Admin will configure settings, manage user accounts, monitor system performance, and

resolve technical issues. They will have the authority to access user details and patient details obtained from hospitals as well.

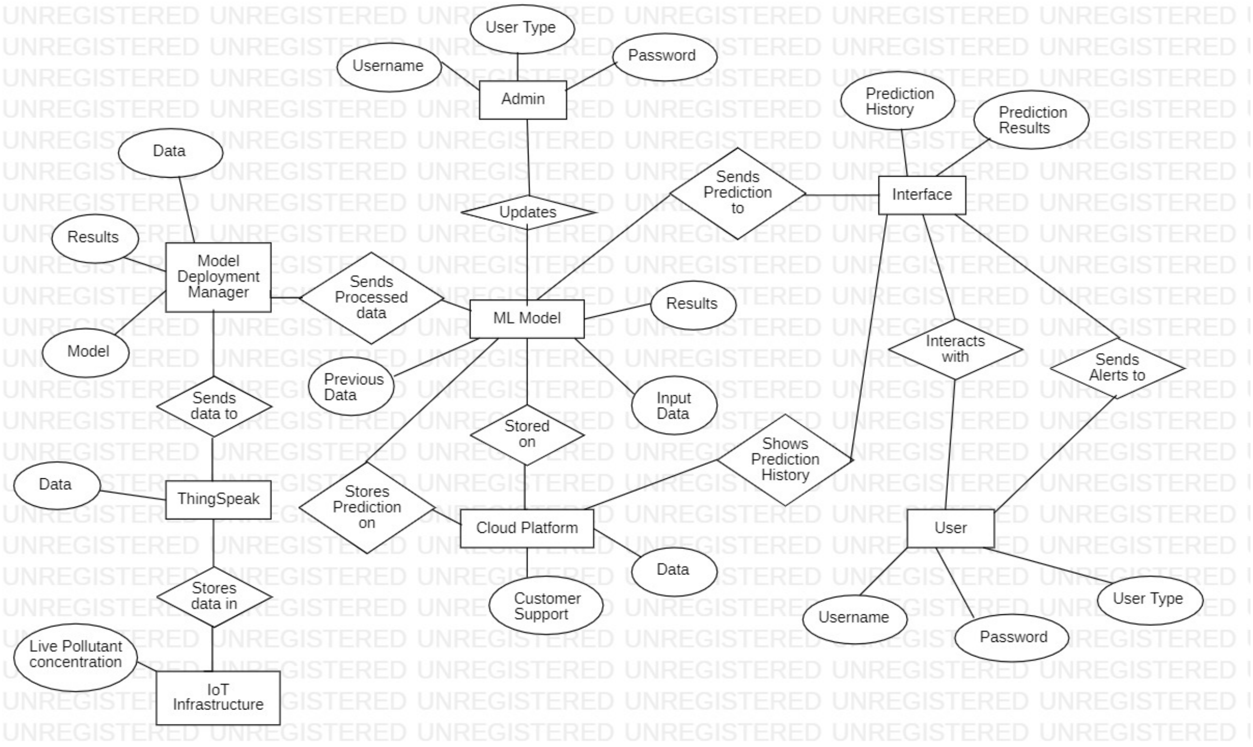
- **Cloud platforms:** Cloud platform refers to a set of cloud-based tools and services that will be used to deploy, manage, and scale applications and services. It will perform tasks such as deploying applications, scaling resources, managing data, and monitoring performance.

  We will make use of two different cloud platforms: Thingspeak to receive data from the sensor stations and to send data to the cloud platform that hosts the model. Streamlit, Spaces or AWS to host the machine learning model.

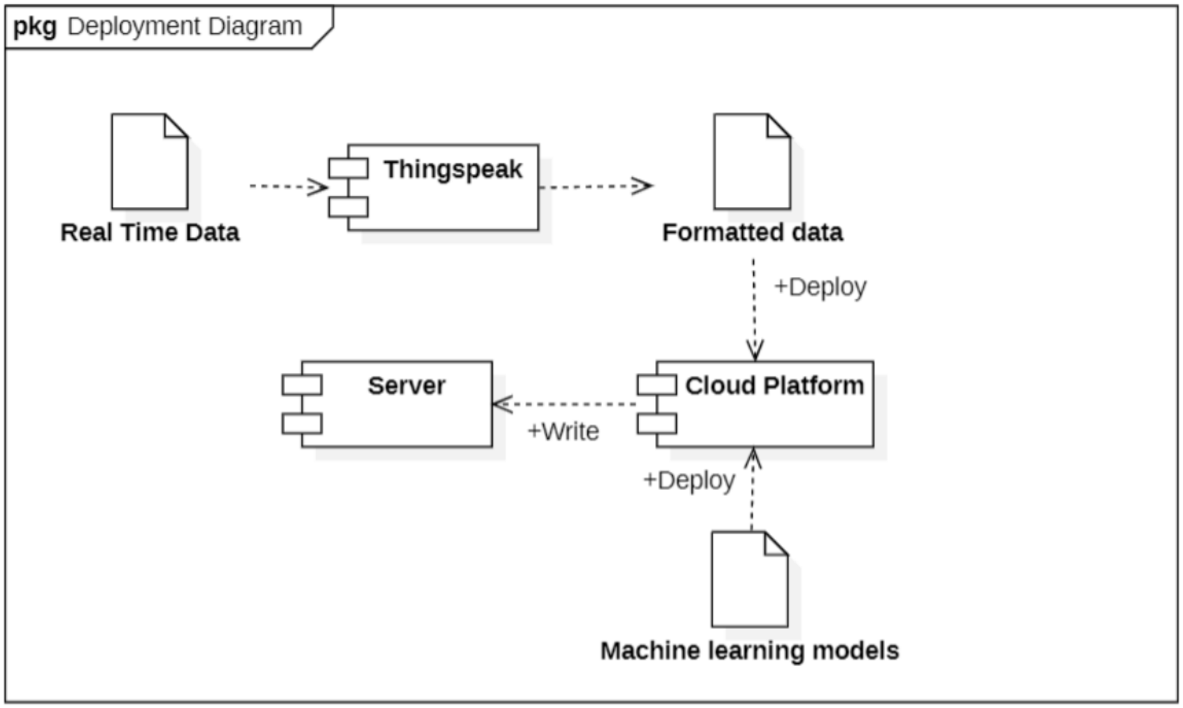
6.3.2 Class Diagram:



6.3.3 ER DIAGRAM:



6.3.4 Deployment Diagram:



6.4 Design Details

6.4.1 Novelty

- The projects on this subject to date have barely researched the consequences of the observed air quality on health, and the predictions are based on outdated and static data.
- We aim to use real time data to make predictions instead of static and outdated data.
- Most of the projects use just the LSTM model, we intend to use an adaptive LSTM that is dynamically capable of adapting to changes.

- Optimization of already existing models for relevant data to produce practical and useful results.

6.4.2 Innovativeness

- Combining three domains- machine learning, cloud computing, and IoT
- Using a hybrid network of forecasting models- ARIMA+Adaptive LSTM
- Increased efficiency to produce an advantageous output.
- By using a cloud platform for all our major functions, we will ensure the entire system is scalable, both horizontally and vertically.
- The system can be integrated into many devices due to the use of a cloud platform.
- The design will be durable, and long-lasting. It will also be relevant because it is malleable enough to accommodate multiple types of parameters.

6.4.3 Reliability

- The project involves complex technical components, such as machine learning algorithms, cloud computing, and IoT devices.
- Any technical issues with these components could affect the accuracy and reliability of the study result.
- We are using tested and industry trusted cloud platforms for the deployment of the system.
- Ensuring consistent performance and functionality across different contexts and use cases.
- Incorporating redundancies and fail-safe mechanisms to prevent system failures.
- We will design the system in modules that can help identify and replace any failed modules, thus improving overall reliability.

- We will conduct extensive testing during the design and development phases to help identify potential problems and improve reliability.
- We will incorporate monitoring and feedback systems that can help detect and diagnose any issues that arise during use, allowing for timely repair or replacement of any faulty components.

6.4.4 Privacy and Security

- The dataset obtained does not contain the patient's personal details, thereby supporting privacy.
- The system should be secure and protect data privacy, as IoT devices are often vulnerable to attacks.
- We will be using a cloud platform to store all the data as well as the model. Cloud platforms offer several security features to ensure that data stored on them is secure.
- Admin access will be given only to trusted individuals, thereby protecting the personal details of users and patients.
- We will only collect and store the minimum amount of data necessary to perform the service. This reduces the amount of data that can be compromised in the event of a security breach.
- We will regularly test the security of the system and update security measures as needed. This can help prevent vulnerabilities from being exploited.
- We will try to use strong encryption methods to protect user data.

6.4.5 Performance

- The system should be able to manage enormous volumes of data and process it in real time.
- The ML model should give an accuracy of 90%+ with a minimum error.

- The use of sensors that are known to give high amounts of accuracy in reasonable time periods will improve the performance of the whole system.
- A hybrid model will make sure to get the best of both worlds, providing us with the best features of both.
- We will efficiently use space, which will lead to a more compact design, which can improve portability, reduce weight, and improve performance.
- We have chosen the most relevant and informative features to avoid overfitting or underfitting the model.
- Tuning hyperparameters has a significant influence on model performance. We will tune the hyperparameters to best fit our model.

6.4.6. Interoperability

- To allow data exchange and analysis, the system should be able to interact with other systems or platforms, such as electronic health records or public health databases.
- The data is collected and monitored on a cloud platform that can be accessed from any device.
- Adopting open standards for data formats, communication protocols, and APIs.
- Clearly defining the interfaces between systems, including the format of data exchanged and the methods of communication.
- Clear and comprehensive documentation.
- Rigorously testing your system for compatibility with other systems can help identify and address interoperability issues before they become a problem.
- Following industry standards for interoperability, such as those defined by standards bodies or industry consortia, can help ensure that your system can work effectively with others in your industry or ecosystem.

6.4.7 Maintainability

- The database and computer maintenance will be handled by the cloud platforms. The workstations must be maintained by the owners.
- The system should be easy to maintain and update, with minimal downtime.
- We will implement a modular architecture that separates different components of the model, such as data processing, feature extraction, model training, and prediction. This allows for easier modification of each component without affecting the entire model.
- We will use version control tools, such as Git, to track changes to the ML model and its codebase. This allows for easy collaboration among team members and helps ensure that changes can be reverted if necessary.
- We will monitor the system regularly to detect any performance degradation or changes in the data distribution. This helps ensure that the model remains accurate and up-to-date.

6.5 Constraints, dependencies, and assumptions:

- **Interoperability requirement**s-

  Basic information can be displayed on an LCD connected to the stations. More data can be viewed on a computer with a basic internet connection. All error messages will be sent to the cloud platform. Some errors can be displayed on the LCD.
- **Interface/protocol requirements-**

  The accuracy of the data collected through IoT devices depends on the quality and reliability of the sensors used. There may also be biases in the data collected, which could affect the validity of the study results. Therefore, maintaining the performance of the IOT sensors is necessary.

- **Discuss the performance related issues as releva**nt-

  The adaptive LSTM model might not have good accuracy if the amount of data fed is limited. The hybrid model might not give good accuracy if the data is irregular and contains anomalies that should be removed.

- **End-user environment-**

  Should provide real-time monitoring, continuously monitor the air quality using IoT devices, and update the database in real-time. The ML models should also be updated periodically to ensure that they remain accurate.

- **Availability of Resources-**

  Finding a skin cancer dataset and a global lung cancer dataset.

- **Hardware or software environment-**

  The amount of energy used by IoT technology is significant, and it needs to be running constantly**.**

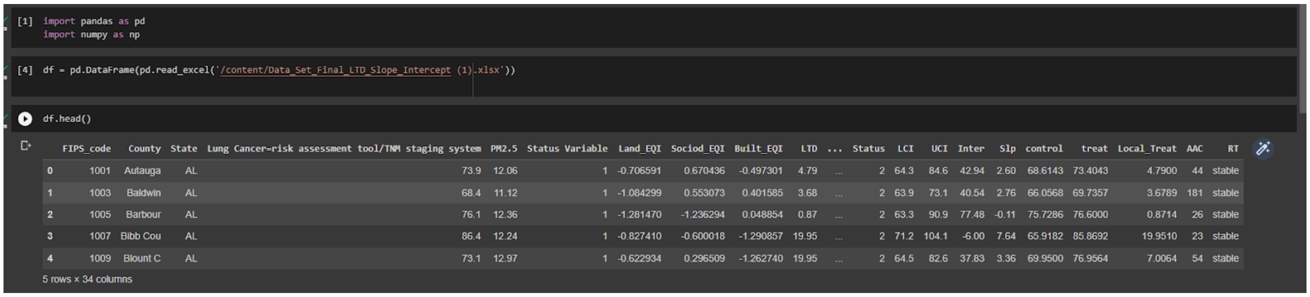
- **Issues related to deployment in target environment, maintainability, scalability, and availability-**

  The data produced by IoT devices is often unstructured and provides a limited perspective due to its massive size.

CHAPTER VII

IMPLEMENTATION AND PSEUDOCODE

7.1. DATA PREPARATION:



```
[1] import pandas as pd
    import numpy as np
```

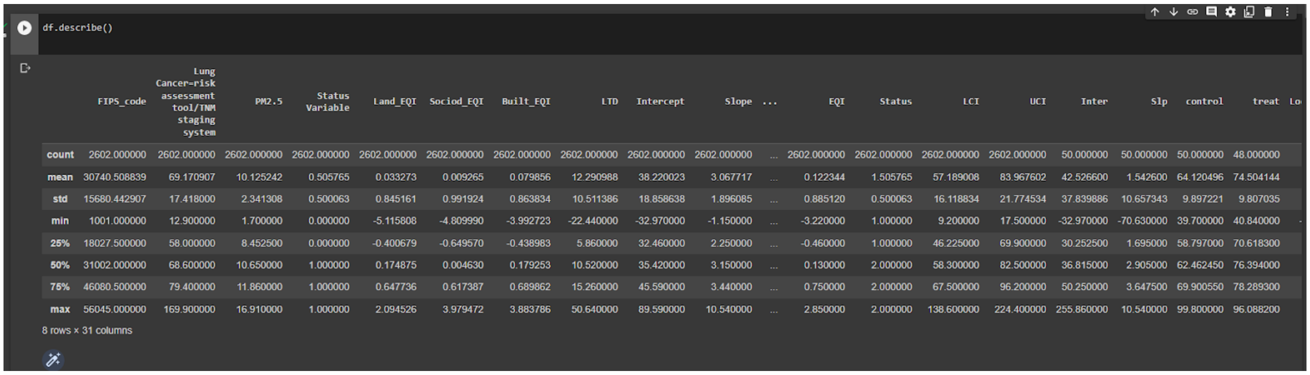
```
[4] df = pd.DataFrame(pd.read_excel('/content/Data_Set_Final_LTD_Slope_Intercept (1).xlsx'))
```

```
df.head()
```

	FIPS_code	County	State	Lung Cancer=risk assessment tool/TNM staging system	PM2.5	Status Variable	Land_EQI	Sociod_EQI	Built_EQI	LTD	...	Status	LCI	UCI	Inter	Slp	control	treat	Local_Treat	AAC	RT
0	1001	Autauga	AL	73.9	12.06	1	-0.706591	0.670436	-0.497301	4.79	...	2	64.3	84.6	42.94	2.60	68.6143	73.4043	4.7900	44	stable
1	1003	Baldwin	AL	68.4	11.12	1	-1.084299	0.553073	0.401585	3.68	...	2	63.9	73.1	40.54	2.76	66.0568	69.7357	3.6789	181	stable
2	1005	Barbour	AL	76.1	12.36	1	-1.281470	-1.236294	0.048854	0.87	...	2	63.3	90.9	77.48	-0.11	75.7286	76.6000	0.8714	26	stable
3	1007	Bibb Cou	AL	86.4	12.24	1	-0.827410	-0.600018	-1.290857	19.95	...	2	71.2	104.1	-6.00	7.64	65.9182	85.8692	19.9510	23	stable
4	1009	Blount C	AL	73.1	12.97	1	-0.622934	0.296509	-1.262740	19.95	...	2	64.5	82.6	37.83	3.36	69.9500	76.9564	7.0064	54	stable

5 rows × 34 columns

Data description:

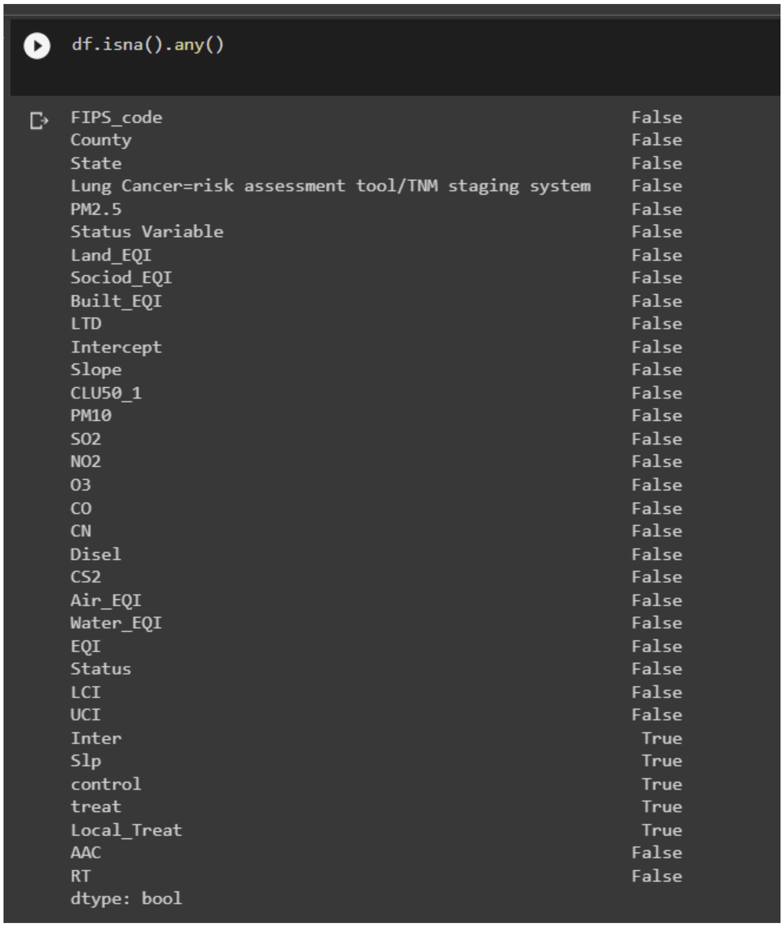


```
df.describe()
```

	FIPS_code	Lung Cancer=risk assessment tool/TNM staging system	PM2.5	Status Variable	Land_EQI	Sociod_EQI	Built_EQI	LTD	Intercept	Slope	...	EQI	Status	LCI	UCI	Inter	Slp	control	treat	Lo
count	2602.000000	2602.000000	2602.000000	2602.000000	2602.000000	2602.000000	2602.000000	2602.000000	2602.000000	2602.000000	...	2602.000000	2602.000000	2602.000000	2602.000000	50.000000	50.000000	50.000000	48.000000	
mean	30740.508839	69.170907	10.125242	0.505765	0.033273	0.009265	0.079856	12.290988	38.220023	3.067717	...	0.122344	1.505765	57.189008	83.967602	42.526600	1.542600	64.120496	74.504144	
std	15680.442907	17.418000	2.341308	0.500063	0.845161	0.991924	0.863834	10.511386	18.858638	1.896085	...	0.885120	0.500063	16.118834	21.774534	37.839886	10.657343	9.897221	9.807035	
min	1001.000000	12.900000	1.700000	0.000000	-5.115808	-4.809990	-3.992723	-22.440000	-32.970000	-1.150000	...	-3.220000	1.000000	9.200000	17.500000	-32.970000	-70.630000	39.700000	40.840000	-
25%	18027.500000	58.000000	8.452500	0.000000	-0.400679	-0.649570	-0.438983	5.860000	32.460000	2.250000	...	-0.460000	1.000000	46.225000	69.900000	30.252500	1.695000	58.797000	70.618300	
50%	31002.000000	68.600000	10.650000	1.000000	0.174875	0.004630	0.179253	10.520000	35.420000	3.150000	...	0.130000	2.000000	58.300000	82.500000	36.815000	2.905000	62.462450	76.394000	
75%	46080.500000	79.400000	11.860000	1.000000	0.647736	0.617387	0.689862	15.260000	45.590000	3.440000	...	0.750000	2.000000	67.500000	96.200000	50.250000	3.647500	69.900550	78.289300	
max	56045.000000	169.900000	16.910000	1.000000	2.094526	3.979472	3.883786	50.640000	89.590000	10.540000	...	2.850000	2.000000	138.600000	224.400000	255.860000	10.540000	99.800000	96.088200	

8 rows × 31 columns

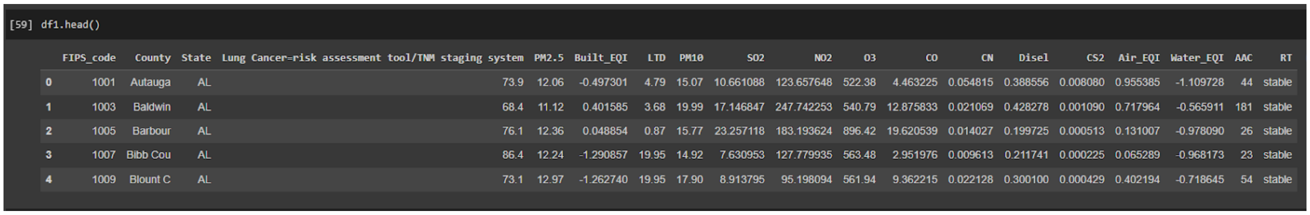
Checking for null values:



```
df.isna().any()
```

```
FIPS_code                                          False
County                                             False
State                                              False
Lung Cancer=risk assessment tool/TNM staging system    False
PM2.5                                              False
Status Variable                                    False
Land_EQI                                           False
Sociod_EQI                                         False
Built_EQI                                          False
LTD                                                False
Intercept                                          False
Slope                                              False
CLU50_1                                            False
PM10                                               False
SO2                                                False
NO2                                                False
O3                                                 False
CO                                                 False
CN                                                 False
Disel                                              False
CS2                                                False
Air_EQI                                            False
Water_EQI                                          False
EQI                                                False
Status                                             False
LCI                                                False
UCI                                                False
Inter                                               True
Slp                                                 True
control                                             True
treat                                               True
Local_Treat                                         True
AAC                                                False
RT                                                 False
dtype: bool
```

Dataset after dropping unnecessary attributes:

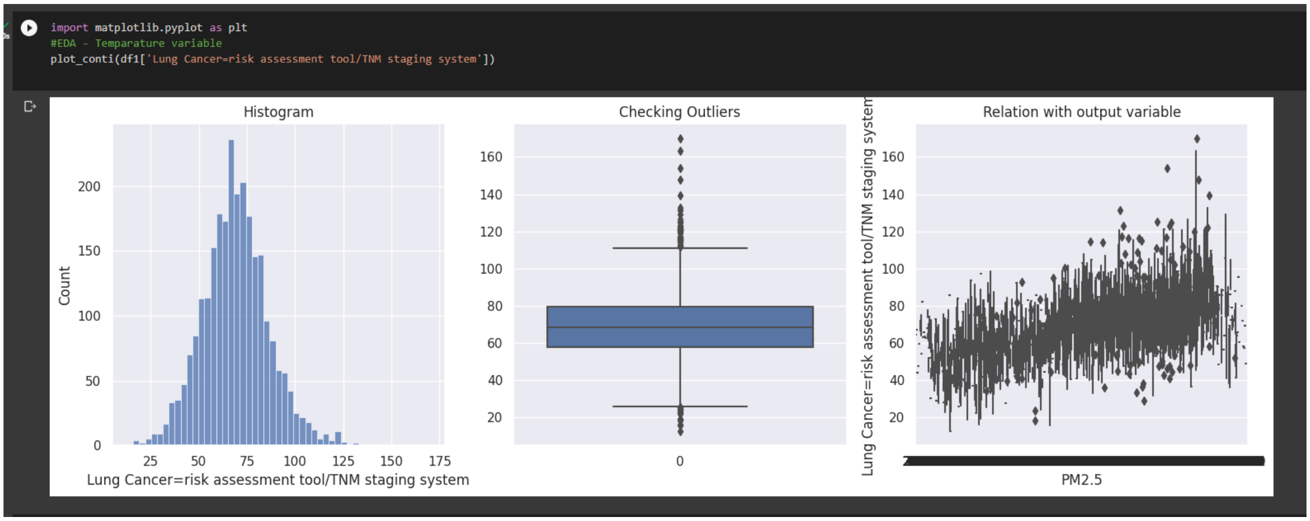


```
[59] df1.head()
```

	FIPS_code	County	State	Lung Cancer=risk assessment tool/TNM staging system	PM2.5	Built_EQI	LTD	PM10	SO2	NO2	O3	CO	CN	Disel	CS2	Air_EQI	Water_EQI	AAC	RT
0	1001	Autauga	AL	73.9	12.06	-0.497301	4.79	15.07	10.661088	123.657648	522.38	4.463225	0.054815	0.388556	0.008080	0.955385	-1.109728	44	stable
1	1003	Baldwin	AL	68.4	11.12	0.401585	3.68	19.99	17.146847	247.742253	540.79	12.875833	0.021069	0.428278	0.001090	0.717964	-0.565911	181	stable
2	1005	Barbour	AL	76.1	12.36	0.048854	0.87	15.77	23.257118	183.193624	896.42	19.620539	0.014027	0.199725	0.000513	0.131007	-0.978090	26	stable
3	1007	Bibb Cou	AL	86.4	12.24	-1.290857	19.95	14.92	7.630953	127.779935	563.48	2.951976	0.009613	0.211741	0.000225	0.065289	-0.968173	23	stable
4	1009	Blount C	AL	73.1	12.97	-1.262740	19.95	17.90	8.913795	95.198094	561.94	9.362215	0.022128	0.300100	0.000429	0.402194	-0.718645	54	stable

7.2. DATA VISUALIZATION:

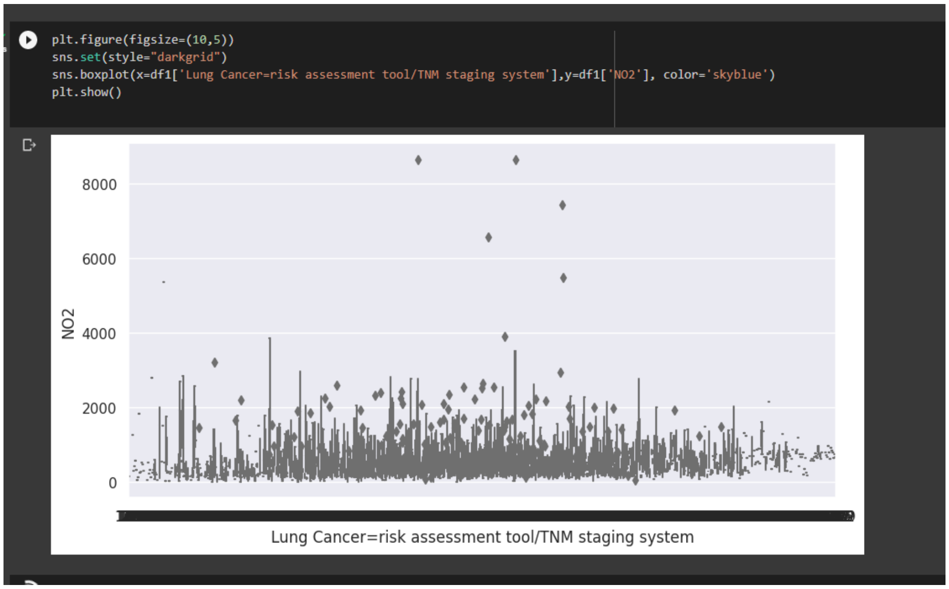
Checking outliers w.r.t lung cancer and PM2.5:



```
import matplotlib.pyplot as plt
#EDA - Temparature variable
plot_conti(df1['Lung Cancer=risk assessment tool/TNM staging system'])
```

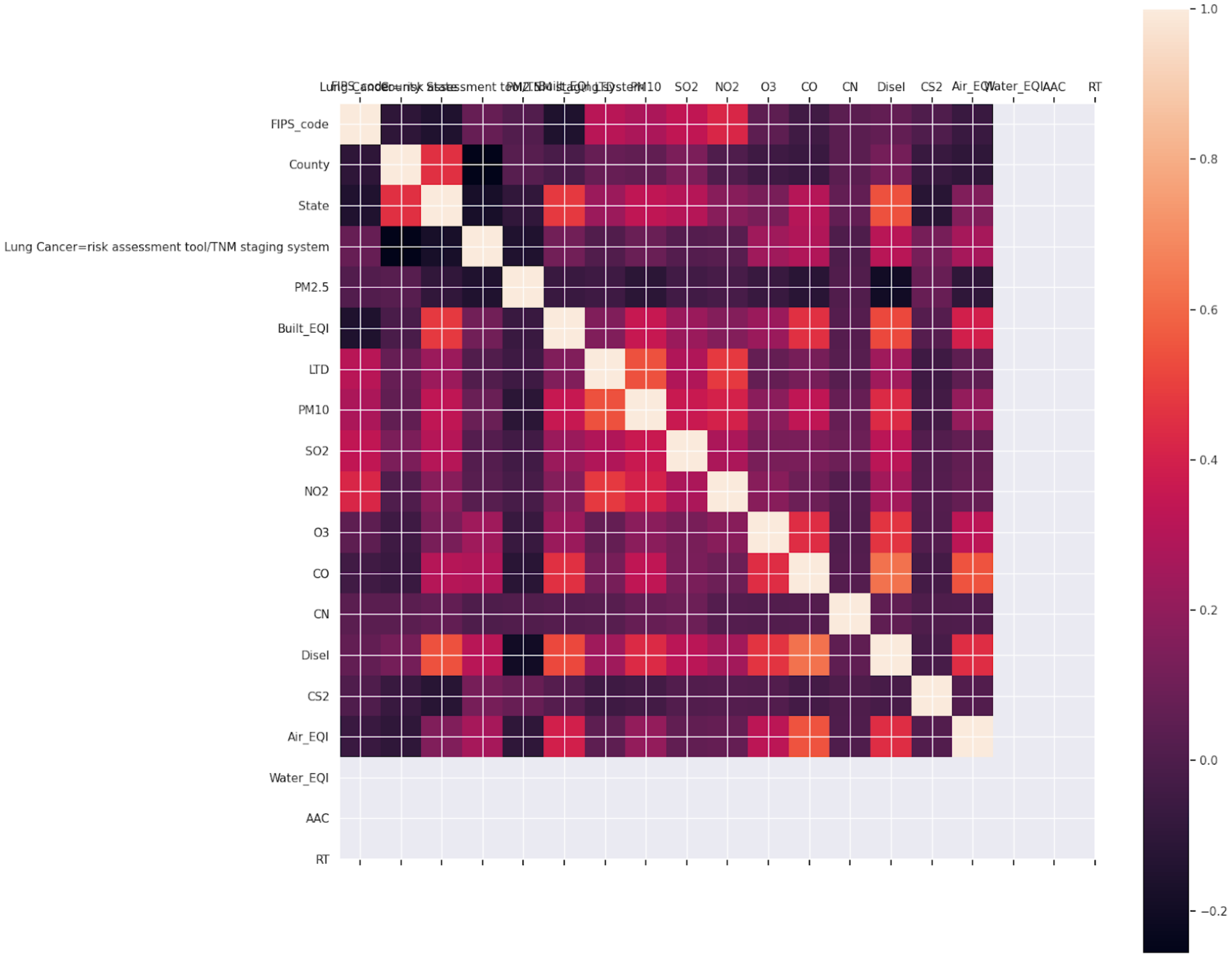
Histogram
Count
0
50
100
150
200
25
50
75
100
125
150
175
Lung Cancer=risk assessment tool/TNM staging system
Checking Outliers
20
40
60
80
100
120
140
160
0
Relation with output variable
Lung Cancer=risk assessment tool/TNM staging system
20
40
60
80
100
120
140
160
PM2.5

Lung cancer vs NO2

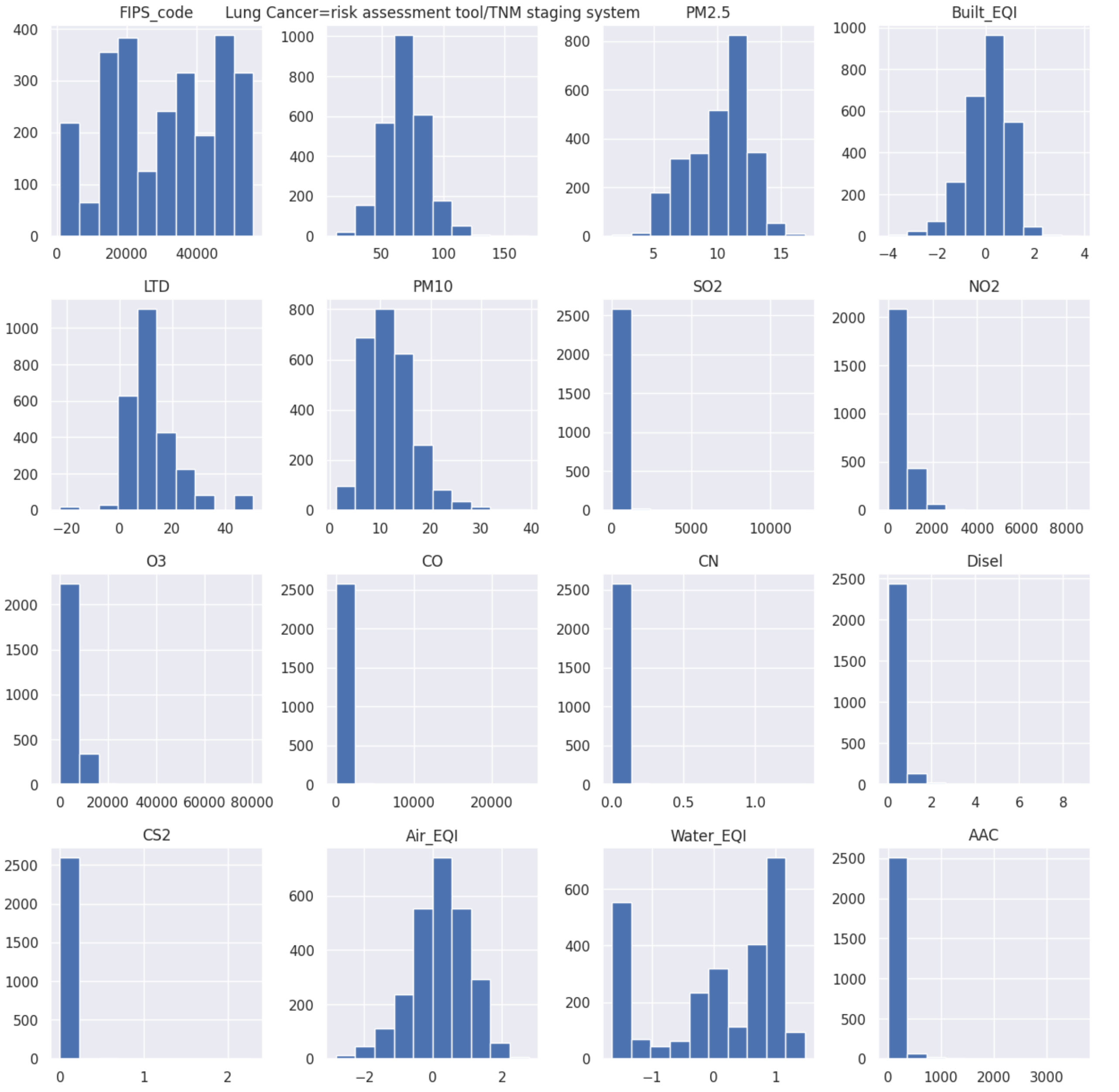


```
plt.figure(figsize=(10,5))
sns.set(style="darkgrid")
sns.boxplot(x=df1['Lung Cancer=risk assessment tool/TNM staging system'],y=df1['NO2'], color='skyblue')
plt.show()
```

Visualizing the correlation between the attributes- CORRELATION MATRIX:



Visualization to check if scaling in necessary:



CHAPTER VIII

CONCLUSION OF CAPSTONE PROJECT PHASE - 1

The industrialization process has contributed significantly to the world's present pollution issue. However, many individuals are still uninformed of the hazards of poor air quality. As a result, it is critical to raise public awareness of these risks and encourage people to take precautions.

After phase-1 of capstone, we have been able to research the feasibility and practicability of achieving our project goals after referencing many academic papers. We have acquired a few datasets, and we already have one large dataset ready for lung cancer.

We have also considered a few standards for combining and extending datasets. We have also looked into a few viable cloud platforms, namely ThingSpeak and Streamlit, but not limited to the quoted options.

We have almost finalized the ML model: bidirectional adaptive LSTM + ARIMA and the hardware we will need for creating the sensor station/IoT infrastructure. We have created architectural and structural diagrams to visualise our proposed system. We will continue expanding the pre-processing steps we have so far implemented.

Naturally, designing this system is not an easy task, but after thorough research, we are now confident in our approach and the direction we are heading in.

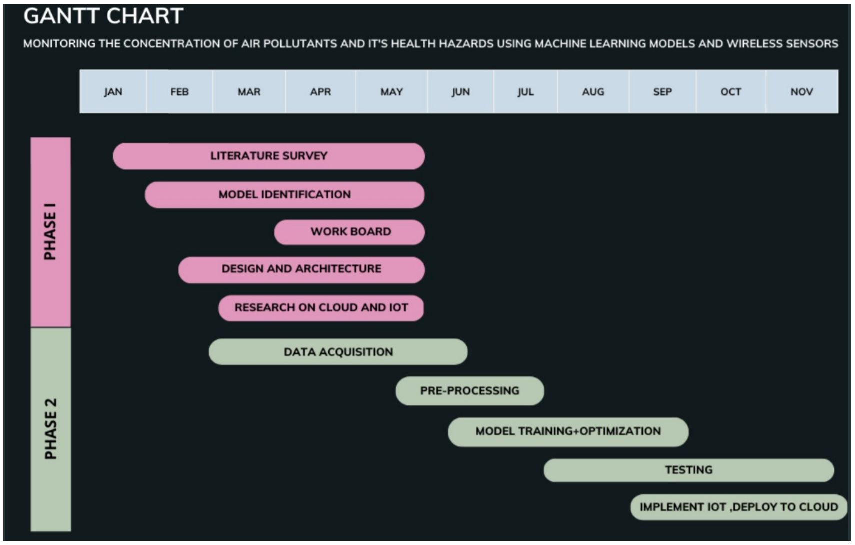
CHAPTER IX

PLAN OF WORK FOR CAPSTONE PROJECT PHASE - 2

We have to now finalize the dataset, model, and cloud platforms being used, which we will use to create our system.

Using this dataset, we will train our hybrid model of Adaptive LSTM and ARIMA. Then we will test the cloud platforms being used and a basic build/prototype of the workstation, followed by the setup of the IoT infrastructure to collect the data.

Throughout our implementation, we will try to firmly stick to the timeline we have decided for our project, which is depicted in the Gantt chart below. We will try to use software engineering principles like Agile methodology to implement our project.



REFERENCES/ BIBLIOGRAPHY

[1] An Application of IoT and Machine Learning to Air Pollution Monitoring in Smart Cities
By: Muhammad Taha Jilani, Husna Gul A.Wahab
[https://ieeexplore.ieee.org/document/8981707]

[2] How Is the Lung Cancer Incidence Rate Associated with Environmental Risks? Machine-Learning-Based Modeling and Benchmarking
By: Kung-Min Wang, Kun-Huang Chen, Shieh-Hsen Tseng
[https://www.mdpi.com/1660-4601/19/14/8445]

[3] Assessment of indoor air quality in academic buildings usng IOT and deep learnings
By: Mohammad Marzouk and Mohammad Atef
[https://www.mdpi.com/1667822]

[4] Household Ventilation May Reduce Effects of Indoor Air Pollutants for Prevention of Lung Cancer: A Case-Control Study in a Chinese Population.
By: Jin Z-Y, Wu M, Han R-Q, Zhang X-F, Wang X-S, et al.
[https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0102685]

[5] Determination of Air Quality Life Index (AQLI) in Medinipur City of West Bengal(India) During 2019 To 2020 : A contextual Study
By: Samiran Rana.
[https://www.researchgate.net/publication/360622768_Determination_of_Air_Quality_Life_Index_Aqli_in_Medinipur_City_of_West_BengalIndia_During_2019_To_2020_A_contextual_Study]

[6] Air pollution and skin diseases: Adverse effects of airborne particulate matter on various skin diseases,

By: Kim Kyung Eun, Cho Daeho, Park Hyun Jeong

[https://pubmed.ncbi.nlm.nih.gov/27018067/]

[7] The spatial association between environmental pollution and long-term cancer mortality in Italy.

By: Roberto Cazzolla Gatti, Arianna Di Paola, Alfonso Monaco, Alena Velichevskaya, Nicola Amoroso, Roberto

[https://www.sciencedirect.com/science/article/pii/S0048969722055383#:~:text=We%20studied%20the%20links%20between%20cancer%20mortality%20and%20environmental%20pollution%20in%20Italy.&text=Tumor%20mortality%20exceeds%20the%20national%20average%20when%20environmental%20pollution%20is%20higher.&text=Air%20quality%20ranks%20first%20for,to%20the%20average%20cancer%20mortality.]

[8] The nexus between COVID-19 deaths, air pollution and economic growth in New York state: Evidence from Deep Machine Learning

By: Cosimo Magazzino , Marco Mele , Samuel Asumadu Sarkodie

[https://www.sciencedirect.com/science/article/pii/S0301479721003030]

APPENDIX A

DEFINITIONS, ACRONYMS AND ABBREVIATIONS

- ARIMA - AutoRegressive Integrated Moving Average.
- LSTM - Long Short-Term Memory
- IoT - Internet of Things
- ML - Machine Learning
- WHO - World Health Organization
- AQI - Air Quality Index
- PM - Particulate Matter
- PMS3003- Plantover Particulate Matter Sensor
- INAAQS - Indian National Ambient Air Quality Standards
- ANN- Artificial Neural Network
- LCD - Liquid Crystal Display
- CNN - Convolutional Neural Network
- LSTM - Long ShortTerm Memory
- IAQ - Indoor Air Quality
- SVM - Support Vector Machine
- IDE - Integrated Development Environment
- OS - Operating System
- AWS - Amazon Web Services
- ROI - Return On Investment

APPENDIX B

SUPPORTING DATA

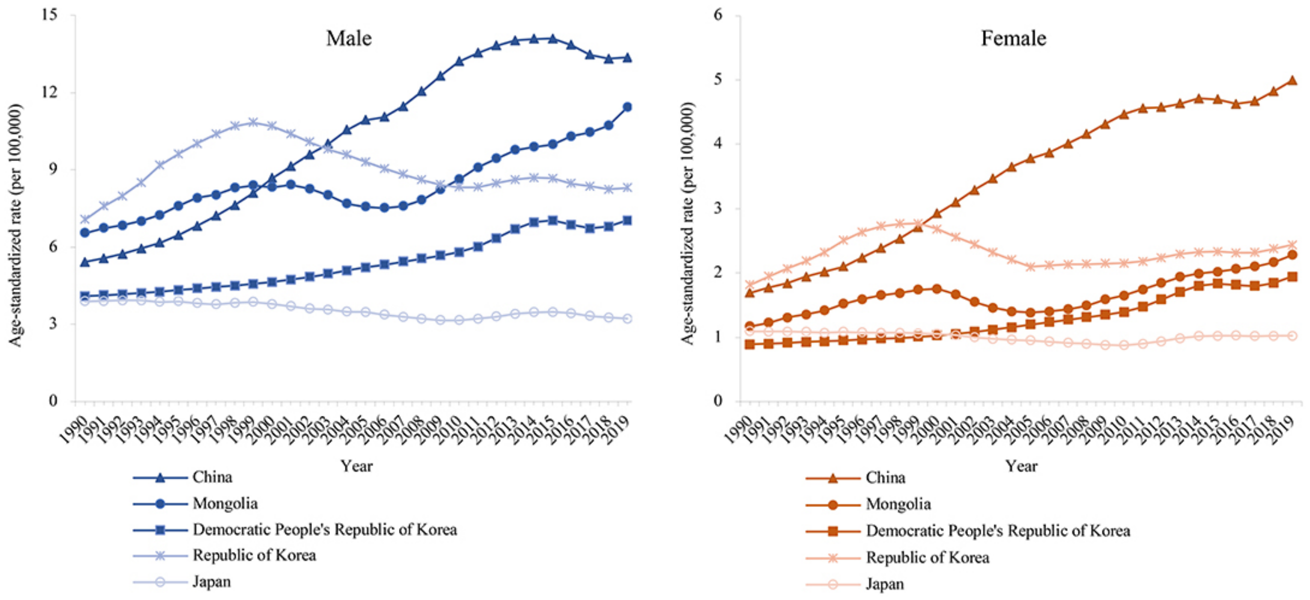


Fig: Lung Cancer death attributable to Long-Term Ambient Particulate Matter (PM2.5)

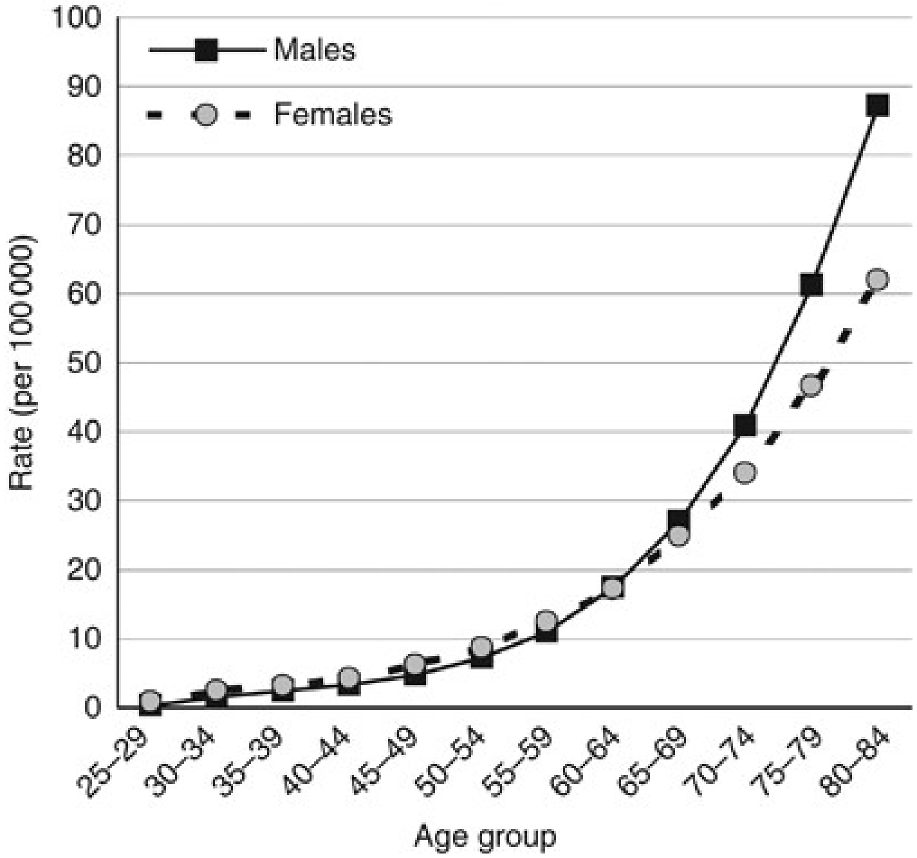


Fig: Tobacco Attributable Cancer trends in Males and Females