UE20CS390A – Capstone Project Review #3

(High Level Design and Proposed Methodology)

Project Title : Monitoring the concentration of air pollutants and its health hazards using Machine Learning models.

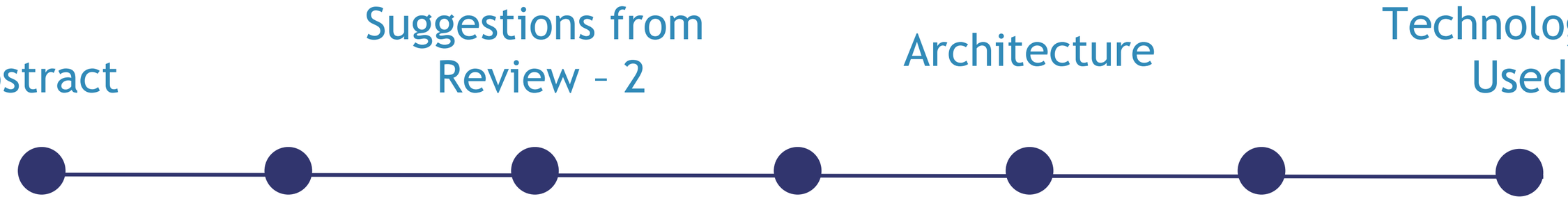
Project ID : 102

Project Guide : Prof. Saritha R

Project Team :

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

Outline


Abstract
Summary of Literature Survey
Suggestions from Review – 2
Proposed Methodology
Architecture
Design Description
Technologies Used

Abstract

Monitoring the concentration of air pollutants & its health hazards using machine learning models

Concerns:

Air pollution has a wide range of negative impacts on the environment and human health like climate change, ozone depletion, acid rain and various harmful diseases.The average deaths per year in India alone is 1.66 million

Problem Statement

We aim to monitor various air pollutants and build a machine learning model to predict the probability of a person contracting various respiratory and skin diseases based on the extent to which they are exposed to harmful pollutants. We will accomplish this by analyzing the pollution data obtained from various wireless sensors and deploy the trained model on cloud

Summary of Literature Survey in Review 2

- The majority of publications cited, take one or a select few risk factors into account. As an illustration, some projects only take into account IAQ or OAQ and not both, or they ignore environmental factors, a person's family history, and occupational exposure, which leads to biased and unreliable conclusions.

- Furthermore, the consequences of the observed air quality on health have barely been researched, and the predictions are based on outdated and static data.

- LSTM and Random Forest were used either as a benchmark, or the main algorithm in many of the research papers, and is proved to be the most effecient algorithms to use.

Suggestions from Review – 2

- Research on the plausibility of correlation between Lung Cancer and COVID-19
- Select datasets from government verified and reliable sources.
- Focus on a small subset of selected diseases like Lung Cancer and Skin Cancer.

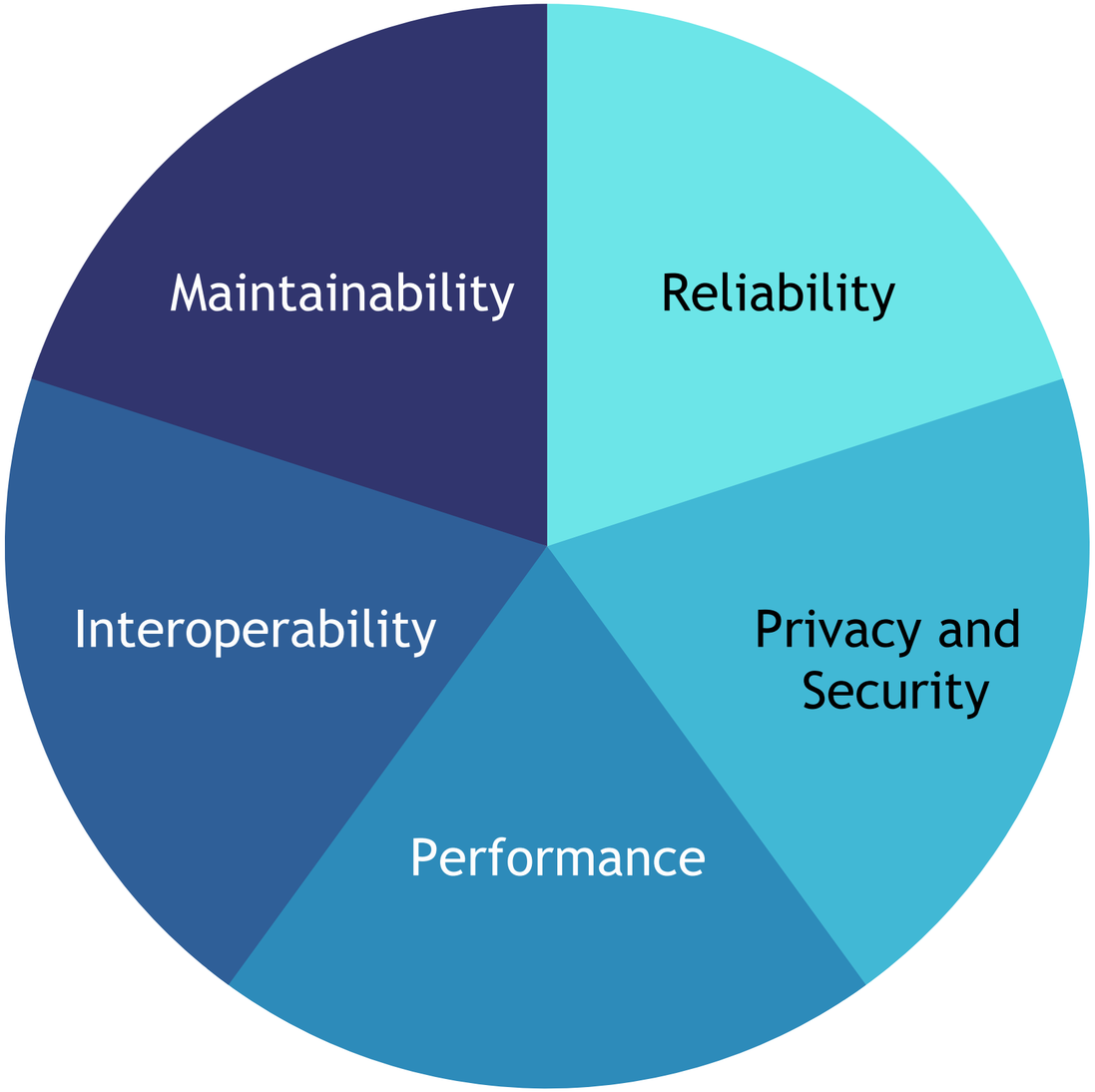
Design Details

- The database and compute 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.

The system should be able to integrate with other systems or platforms, such as electronic health records or public health databases, to facilitate data sharing and analysis.

- The data is collected and monitored on a cloud platform which can be accessed on any device

- The system should be able to handle large amounts of data and provide real-time processing of data
- The ML model should give an accuracy of 90%+ with minimum error



- 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 results

- 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.

Framework Used

Hybrid network of

Adaptive LSTM
(Long-Short-Term-Memory)

ARIMA
(Auto-Regressive-Integrated-Moving-Average)

Why Adaptive LSTM+ARIMA?

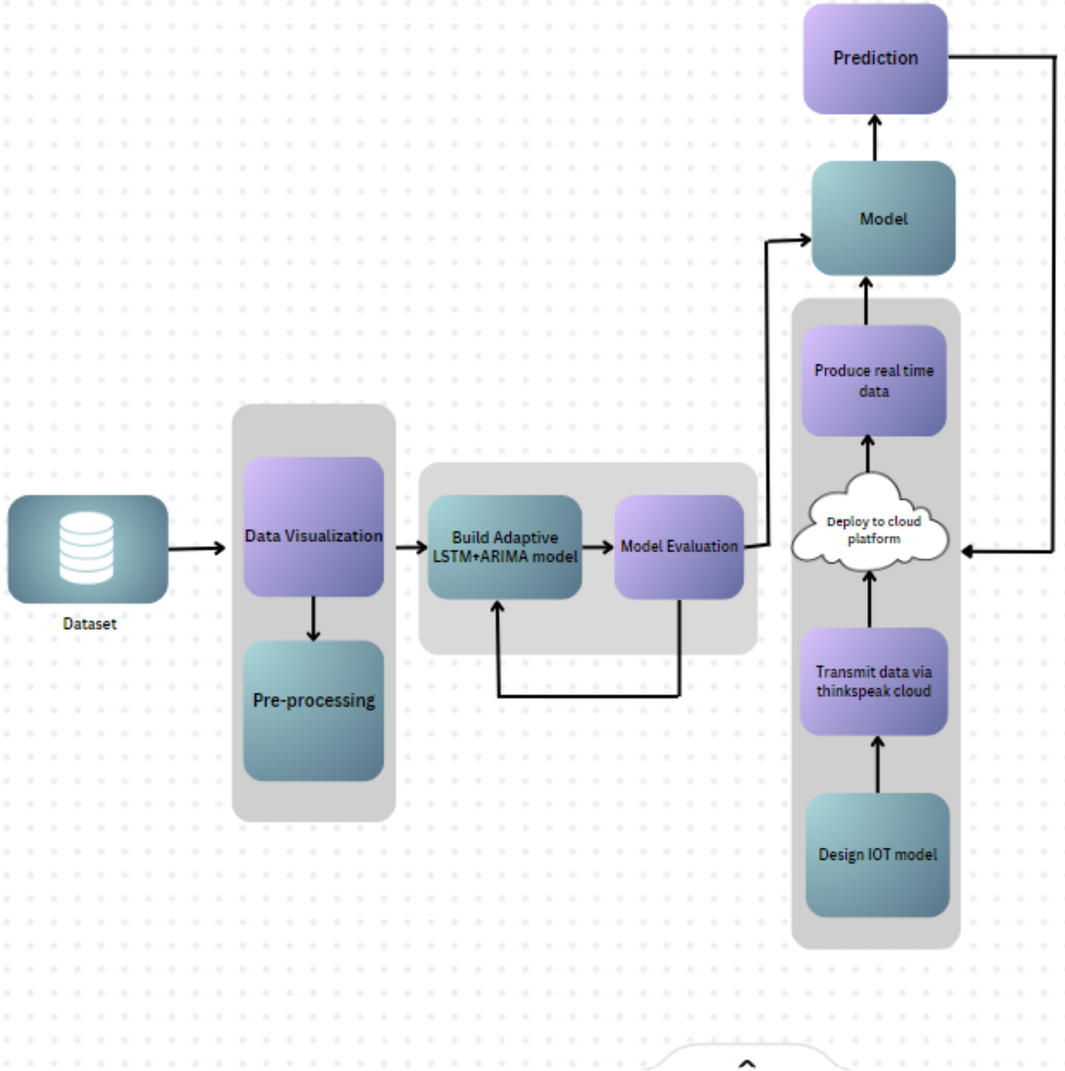
ARIMA is easy to intepret and reliable for small and stationary datasets.

While Adaptive LSTM works best for large, real time,varying lengths, non-stationary data and learns complex patterns.

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.

Architecture Diagram



Workflow

Pre-process your data:
Clean and preprocess your time series data, handling missing values, transforming variables, splitting into testing and training data.

Train ARIMA model:
Fit an ARIMA model - capture the seasonal and trend components of the time series. Tune the ARIMA hyperparameters

Generate ARIMA forecasts:
Use the trained ARIMA model to generate short-term forecasts on the testing data.

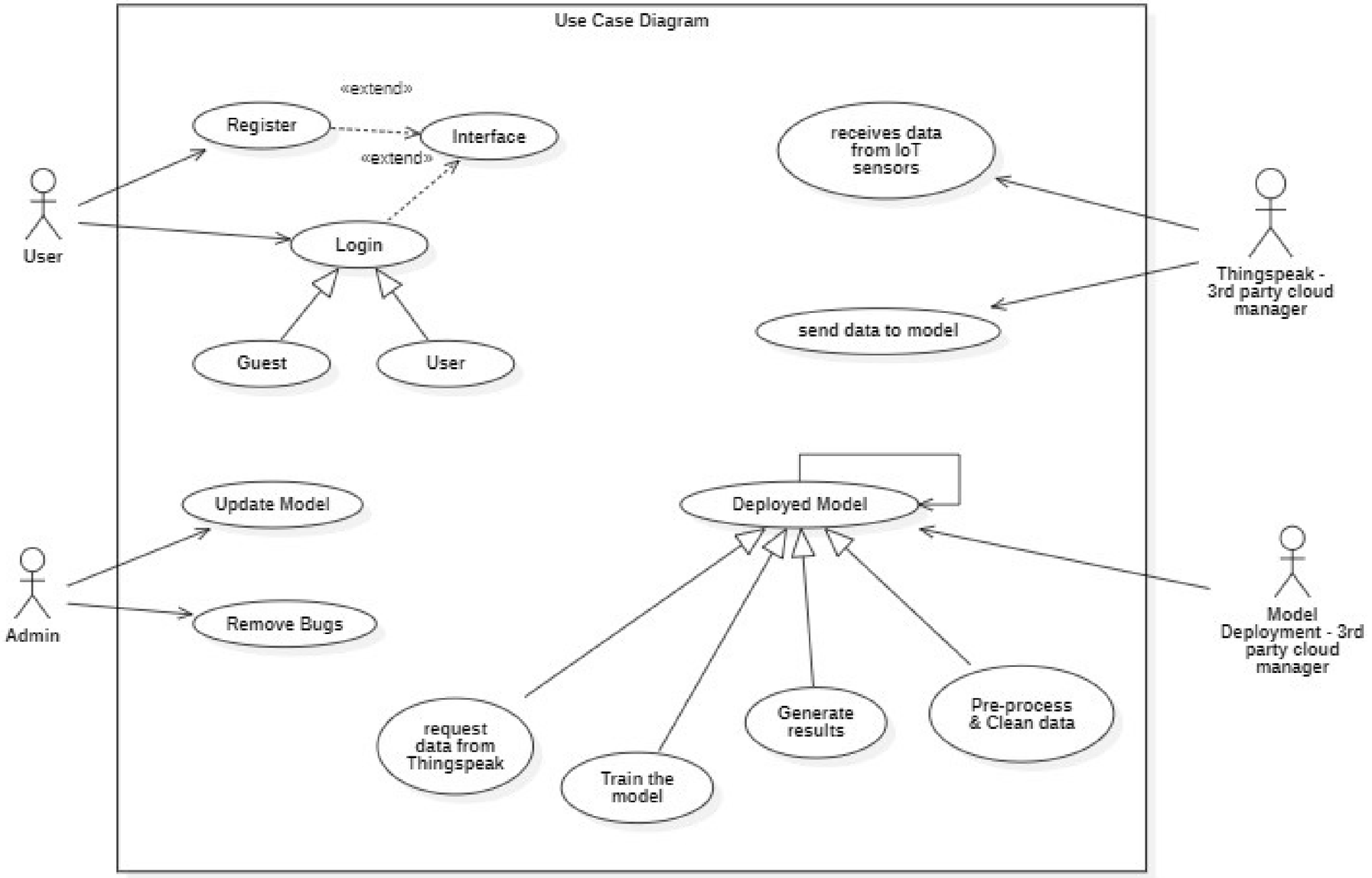
Compute ARIMA residuals:
Calculate the residuals,
for the training and testing data.

Train Adaptive LSTM model:
Use the ARIMA residuals as input to train an Adaptive LSTM model on the training data.

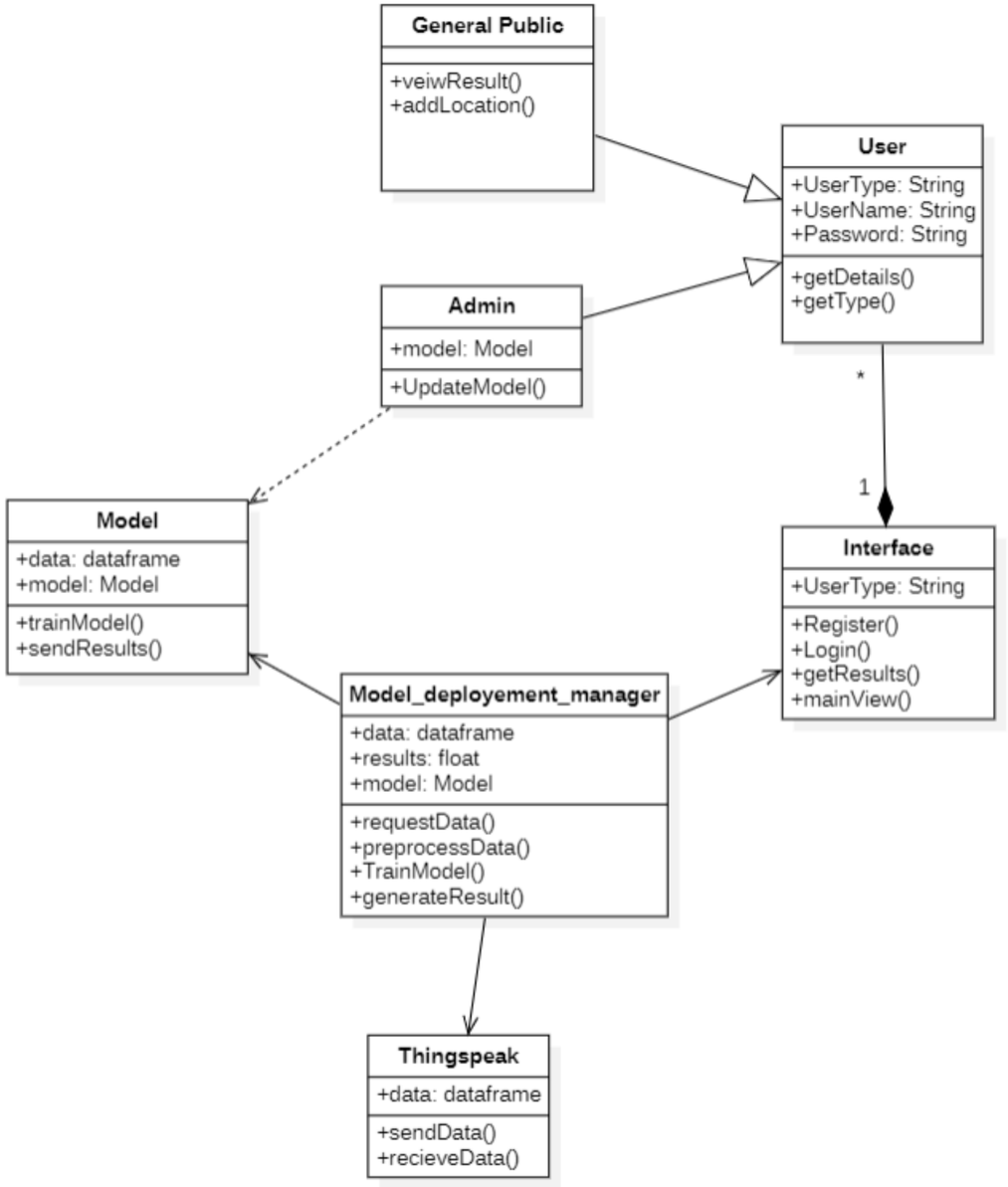
Combine forecasts:
Generate forecasts using the trained Adaptive LSTM model on the testing data and combine them with the ARIMA forecasts obtaining the final hybrid forecast.

Evaluate and optimize:
Evaluate the performance of the hybrid ARIMA-Adaptive LSTM model on the testing data - fine-tune the model as needed.

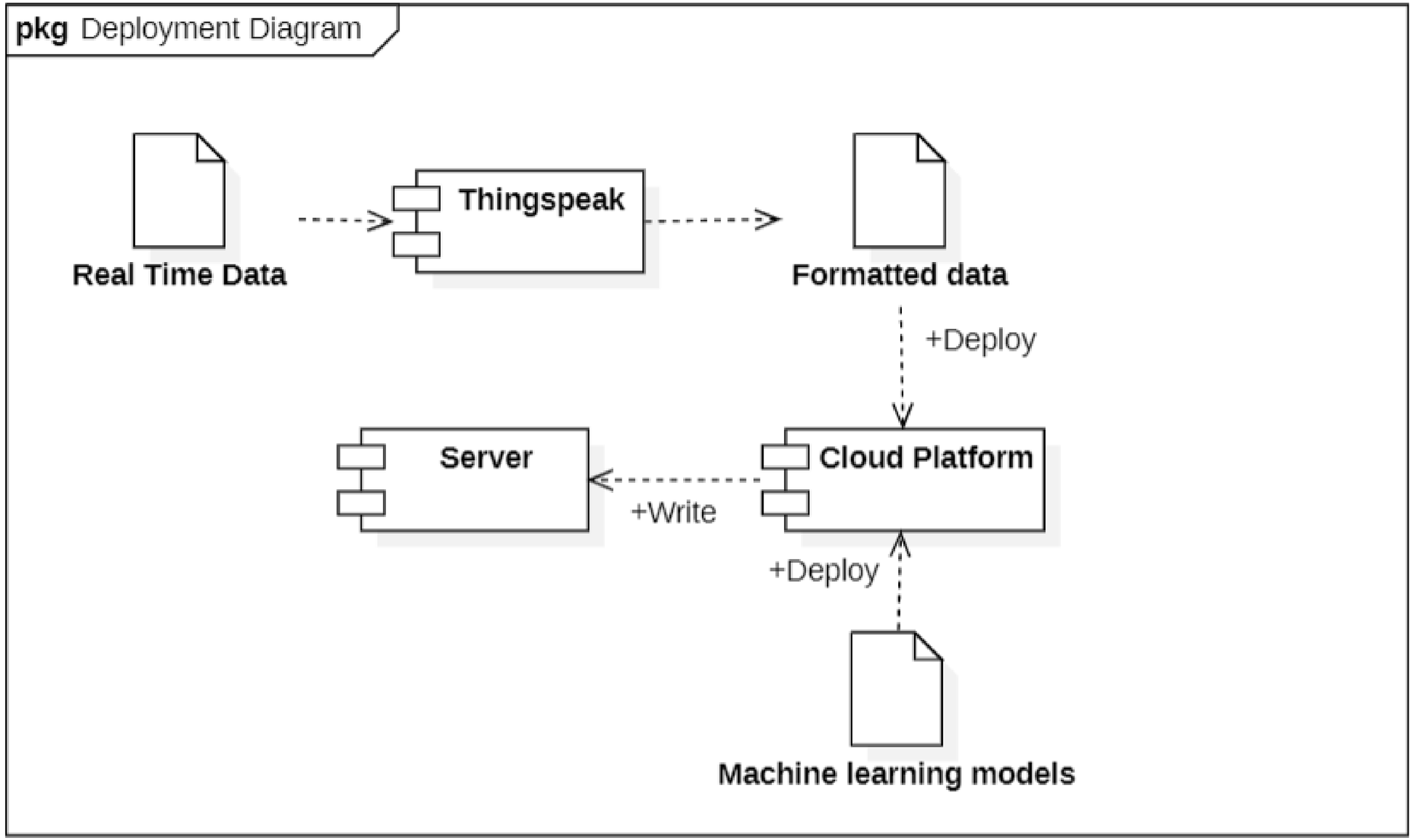
Use-Case Diagram



Class Diagram



Deployment Diagram



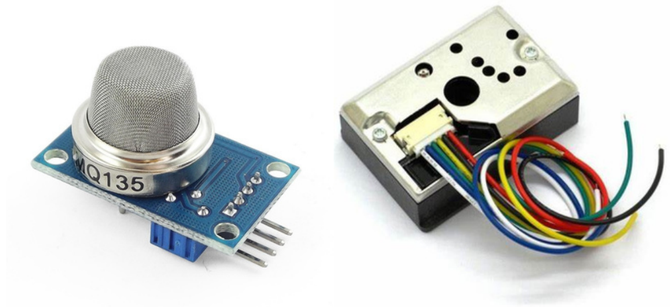
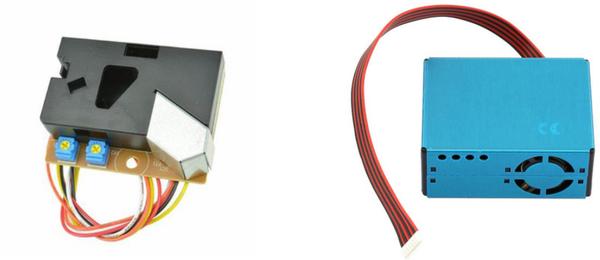
Technologies Used

Languages	Dev Tools	Web Resources
Python and C++ will be used in this project. Python will be used for our model C++ will be used to code the IOT sensors and data transmission	IDEs and code editors like VS Code and Arduino IDE will be used to code. Various modules and libraries will be used in creating the model and for the standard protocols for data transmission.	We will utilize various resources like a vast amount of research papers, lastminiutenginners, geeksforgeeks, video tutorials etc.

Technologies Used

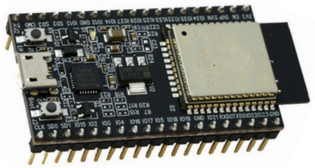
Streamlit CLoud	HuggingFace / Spaces	AWS
Streamlit Pros: • Free hosting, 1GB Storage. • Frontend is easy to make and connect. • Good support from the devs. • Easy to use cloud functionality. Cons: • Relatively new • Less feature rich due to its age • Small community • Things change often	Spaces hf.co/spaces Pros: • Free hosting and storage • Feature rich. • Links up to other platforms well. • Locks you into a few things (Selective GitHub Repos) Cons: • Reliability is sometimes bad. • Little convoluted to use. • The interface and platforms does not inspire stability.	aws Pros: • Extremely feature rich. • Greater flexibility and scalability. • Great control. Cons: • Can be more complicated and difficult to set up and manage. • Can be more expensive. • Requires more expertise and resources to maintain and update.

Technologies Used

Sensors	Sensors	ThingSpeak
		ThingSpeak
MQ 135 will be used to detect CO, CO2, and NH4. GP2Y1010AU0F will be used for dust detection.	DSM501A Dust Sensor will be used to detect dust, PM 2.5 and PM 10 PMS5003 will be used to detect PM 1.0, PM 2.5 and PM 10	We will use ThingSpeak to receive data from the sensor stations and to send data to the cloud platform that hosts the model. Both transmission and receiving data is done through basic API calls.

Technologies Used

ESP32 / ESP 8266



We will use these 2 ESP modules to construct the IOT sensor stations. Both of them are capable of connecting to the internet.

Project Progress

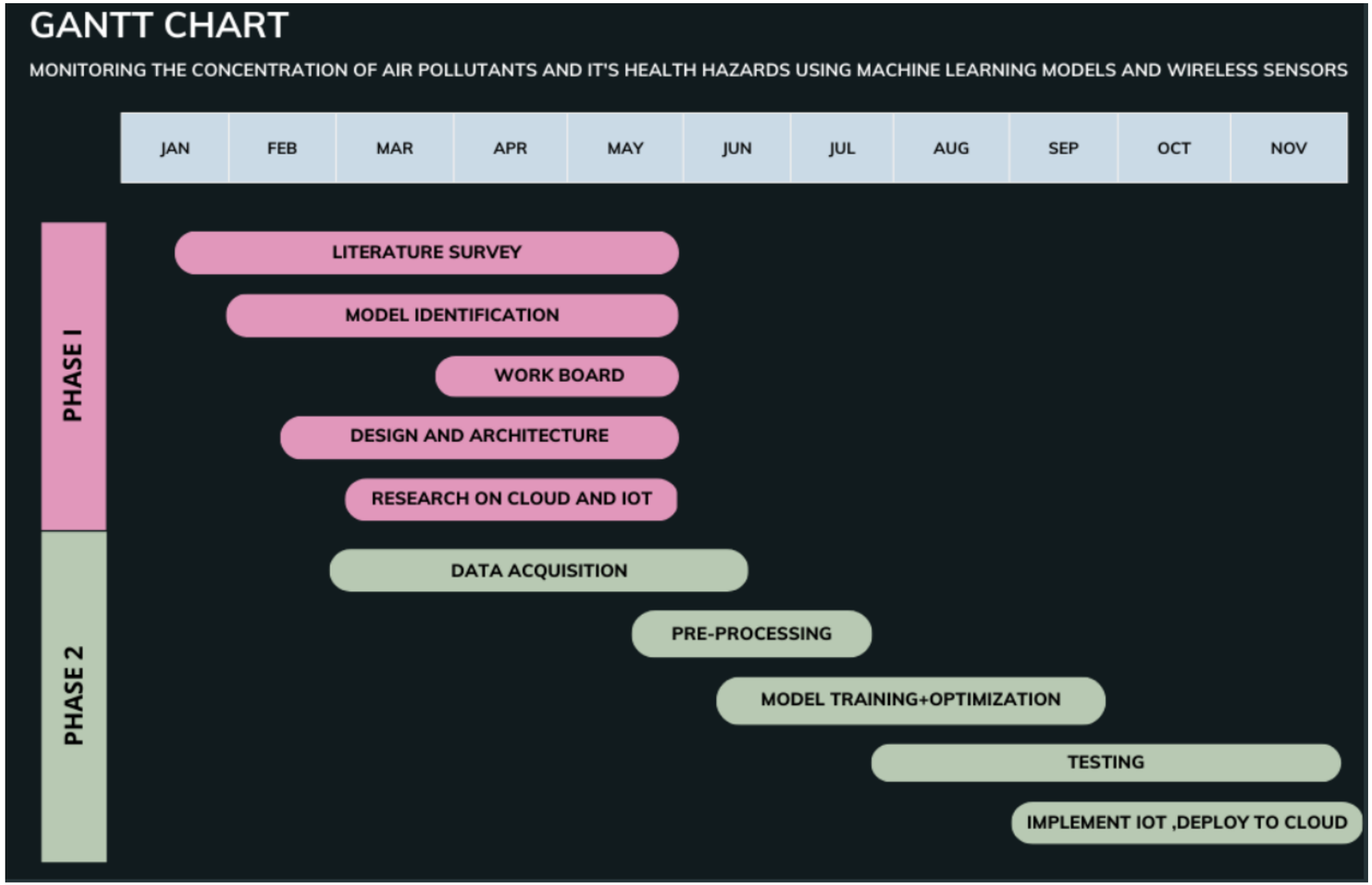
Completed

- Performed literature review.
- Acquired a few datasets.
  - We have one large dataset ready for lung cancer.
- Understood a few standards defined for combining and extending datasets.
- We have looked into a few viable cloud platforms.
  - ThingSpeak
  - Streamlit
- We have almost finalized the ML model.
  - bidirectional adaptive LSTM + ARIMA
- We have created architectural and structural diagrams.

To- Do

- Finalization of the dataset and cloud platforms being used which will promptly be completed before ESA.
- Testing of the cloud platforms being used and a basic build/ prototype of the workstation.
- Setup of the IoT infrastructure to collect the data
- Implementation

Capstone (Phase-I & Phase-II) Project Timeline



References

[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]

References

[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] 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]

Thank You