UE20CS390A – Project Phase – 1
End Semester Assessment

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

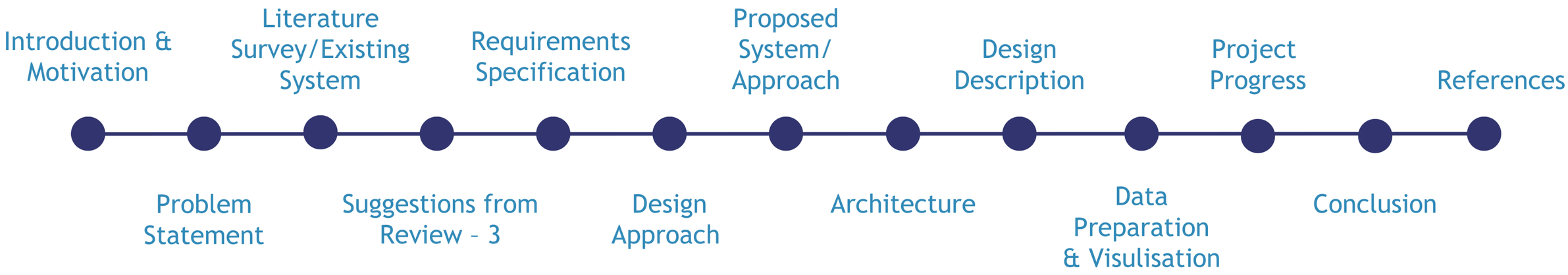
Project ID : 102

Project Guide : Prof. Saritha

Project Team :

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

OUTLINE


Introduction & Motivation
Problem Statement
Literature Survey/Existing System
Suggestions from Review – 3
Requirements Specification
Design Approach
Proposed System/ Approach
Architecture
Design Description
Data Preparation & Visulisation
Project Progress
Conclusion
References

MOTIVATION

- Increase in urbanization and industrialization- rise in air pollution
- Causes over 7 million deaths each year - Lung Cancer - 2 million
- Many people still remain unaware of the dangers of pollutants - **PM2.5, PM10, NO2, and CO**
- Necessary to develop and implement air quality monitoring systems
- Provide individuals with real time information on air quality levels

PROBLEM STATEMENT

Monitoring the concentration of air pollutants like PM2.5, PM10, NO2 and CO and its health hazards (lung cancer) using hybrid model of ARIMA and adaptive LSTM

The proposal is a continuous air quality monitoring system that will keep track of the air quality in the user's vicinity to predict the probability and alert them in case of increased likelihood of developing lung cancer. This will be achieved using a hybrid model of Adaptive LSTM and ARIMA ML models. The model will be deployed on a cloud platform.

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 air in their surroundings.

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 using IoT sensors.

LITERATURE SURVEY

Paper 1

Paper Details	Objective of paper, Techniques/Methods	Advantages	Limitations
Mohammad Marzouk and Mohammad Atef **"Assessment of indoor air quality in academic buildings using IOT and deep learnings"**: Mdpi(June 2022)	To find the correlation between outdoor pollution and indoor air quality, by monitoring real time IAQ using IOT sensors and sends the collected data to cloud via wireless connections.CNN and LSTM are used to train the collected data.	◦ Uses real time data. ◦ Finds the correlation between outdoor and indoor environment. ◦ Concentrates on rural areas too, which might be affected by burning wood, construction particles and unpaved roads. ◦ Achieved more than 90% accuracy for the predicted data.	◦ 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

Paper 2

Paper Details	Objective of paper, Techniques/Methods	Advantages	Limitations
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 National Taiwan University of Science and Technology DOI: 10.3390/ijerph19148445	The objective of the paper is to investigate the relationship between lung cancer incidence rate and air pollution using machine-learning-based modeling and benchmarking. The study aims to develop a predictive model to understand the impact of air pollutants on the disease. They make use of several ML models like Logistic Regression, Random Forest, SVM and Gradient Boosting	◦ They perform benchmarking, i.e, comparing the performance of different ML models. The results showed that the random forest was the best with (RMSE: 4.837, R-squared: 0.895). ◦ The dataset they have used contains both environmental risk factors (air pollutants) and the lung cancer incident rates from various countries and hence increases generalizability	◦ They have used a vast dataset and hence the completeness and accuracy of the data is a challenge. ◦ The paper doesn't consider several other factors like genetics and lifestyle factors that can also impact lung cancer incidence rates. ◦ They do not show casuality. i.e, the study doesn't definitively show that air pollution causes lung cancer as other factors are not considered.

Paper 3

Paper Details	Objective of paper, Techniques/Methods	Advantages	Limitations
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 Department of Political Sciences, Roma Tre University, Italy Department of Political Sciences, University of Teramo, Italy Nord University Business School, Norway https://doi.org/10.1016/j.jenvman.2021.112241	This paper aims to prove a link between various air pollutants (like PM or NO2) and death. Considering the paper is from a department of business they also try to show a link to economic growth. Further it shows a link between unsustainable growth and air pollutants.	◦ Considers individual pollutants. ◦ Very well thought out data gathering and cleaning. ◦ Techniques/Models used: ANN, Deep learning - Oryx Protocol, "Multiple regression" <= derived from another paper, a specialized decision tree. ◦ The above are well suited for the type of predictions and the type of data we are using.	◦ 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.

Paper 4

Paper Details	Objective of paper, Techniques/Methods	Advantages	Limitations
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	This paper analyzed the links between cancer mortality, socio-economic factors, and sources of environmental pollution in Italy, both at wider regional and finer provincial scales, with an artificial intelligence approach. Random Forest (RF) regression coupled with a Boruta feature importance analysis, K means clustering SMR forecasting and Feature importance, Regional cluster analysis	◦ 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. ◦ Explored the potential spatial association between socioeconomic and lifestyle factors	◦ Some punctual sources of pollution do not show a relation with any specific cancer type ◦ This study is local to Italy, and needs to be extended to other countries as well since every area has differnt levels of air poulltion.

FINDINGS

- 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

- Update problem statement to better reflect the scope and nature of our project.
- Find more spread out datasets. We should look for datasets from other areas of the world and focus on getting it near or in India. This will help us in various ways:
  - Firstly it will help us build a more accurate and more flexible and generalized model.
  - Furthermore it will ensure that the model is accurate for use near and inside India.

REQUIREMENTS SPECIFICATIONS

Framework Used

Hybrid network of

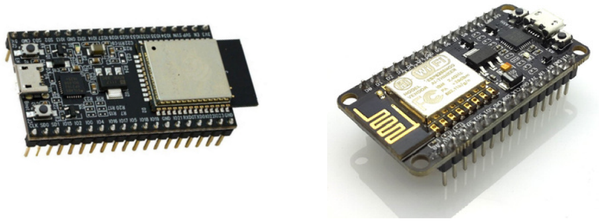
Adaptive LSTM
(Long-Short-Term-Memory)

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

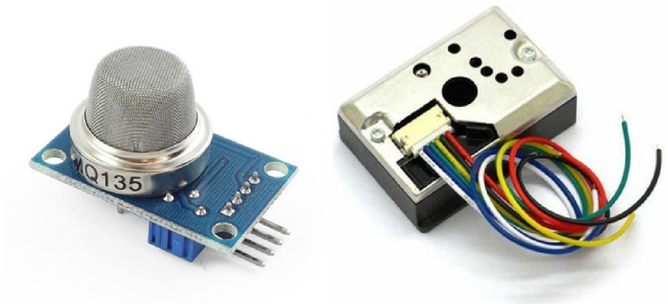
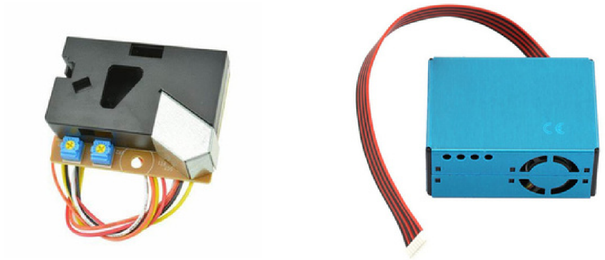
TECHNOLOGIES USED

Languages	Dev Tools	HuggingFace / Spaces
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.	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.

TECHNOLOGIES USED

Streamlit CLoud	AWS	ESP32 / ESP 8266
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 ◦	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.	 These 2 ESP modules to construct the IOT sensor stations. Both of them are capable of connecting to the internet.

TECHNOLOGIES USED

Sensors	Sensors	ThingSpeak
		
MQ 135 will be used to detect CO, NO2, 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	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.

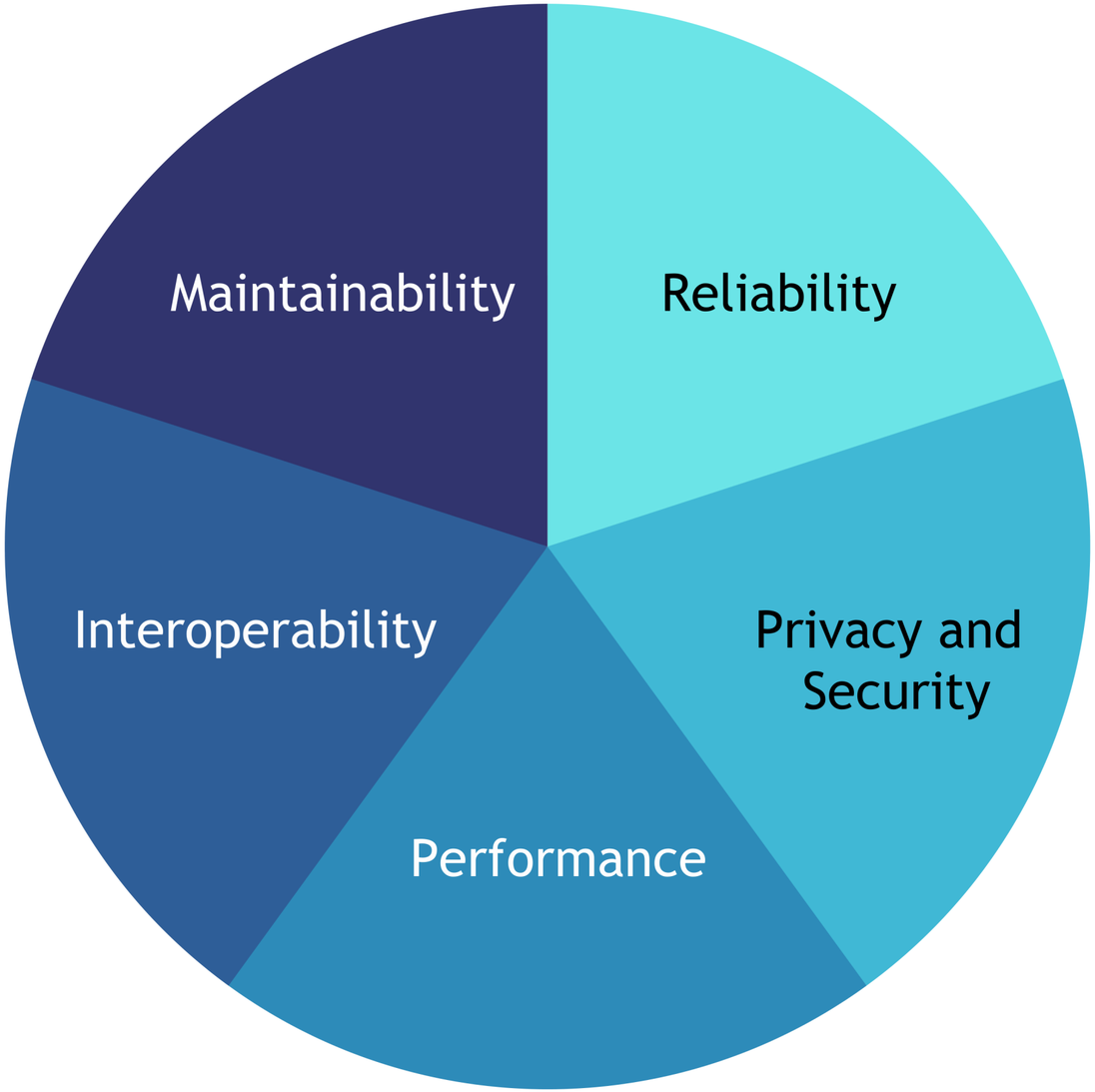
DESIGN DETAILS

- 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

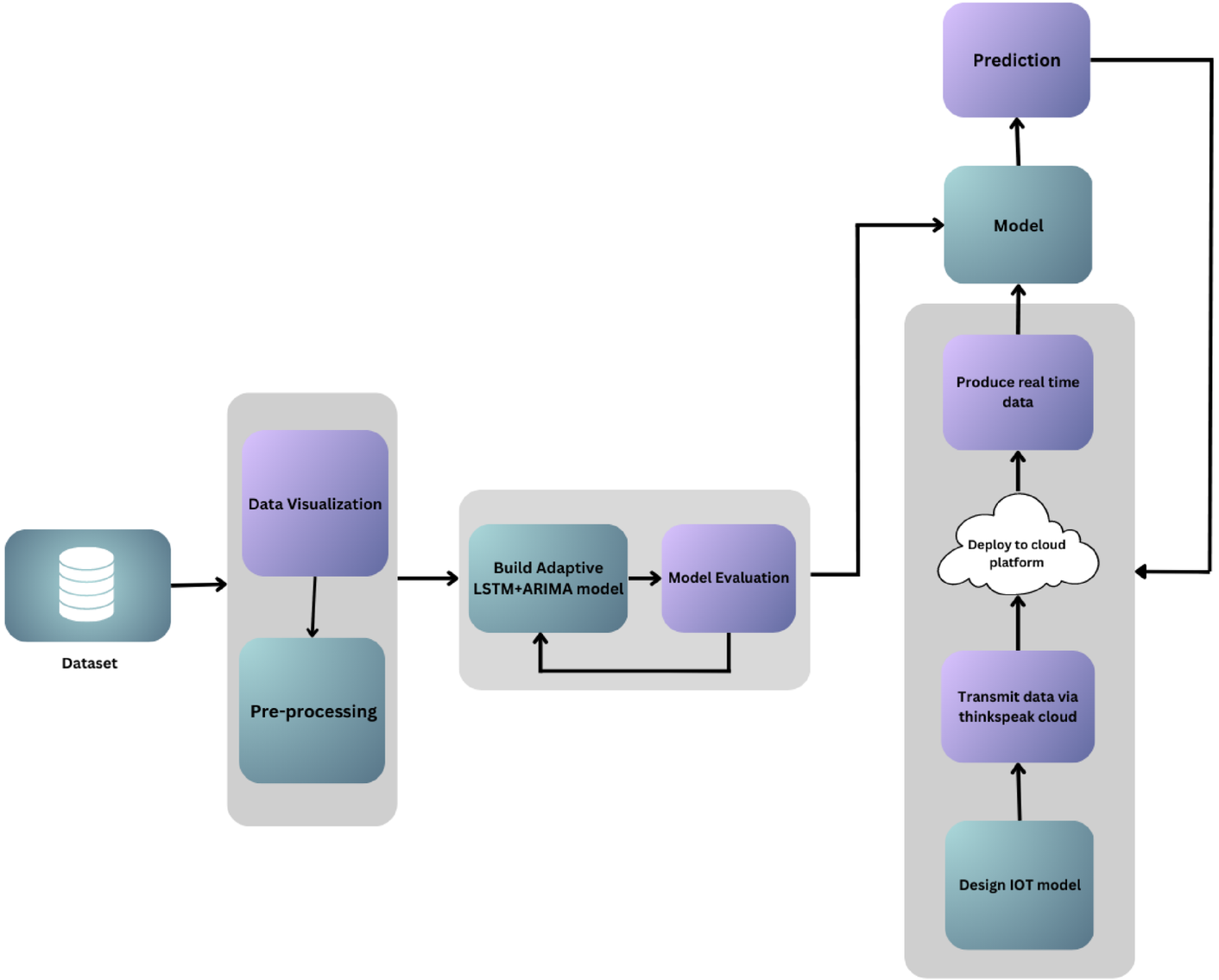
- The system should be able to handle large amounts of data and provide real-time processing of data
- The ML model should give an accuracy of 90%+ with minimum error



- 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.
- 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.

ARCHITECTURE DIAGRAM



WORKFLOW

PES UNIVERSITY

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.

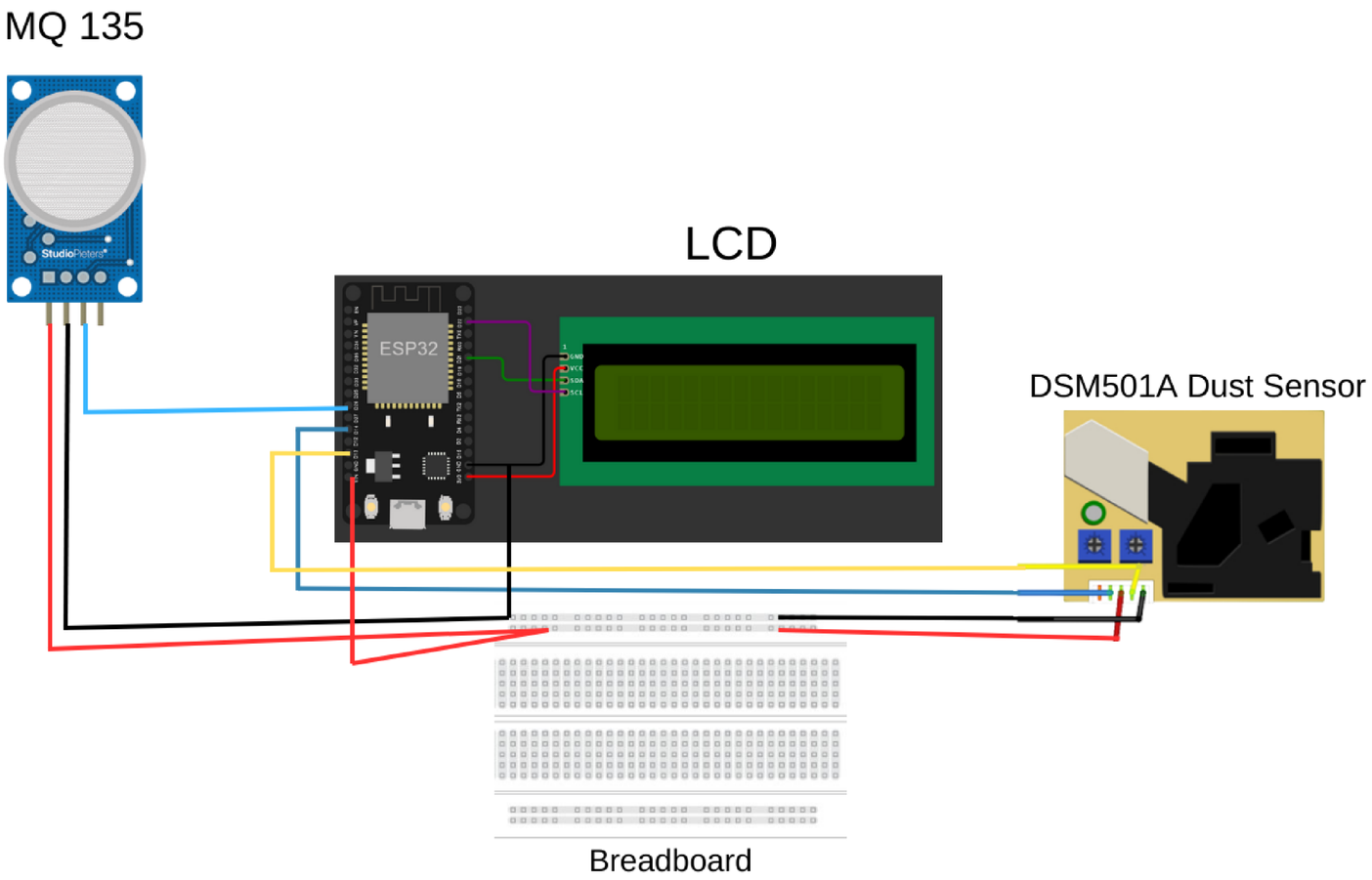
Compute ARIMA residuals:
Calculate the residuals,
for the training and testing data.

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.

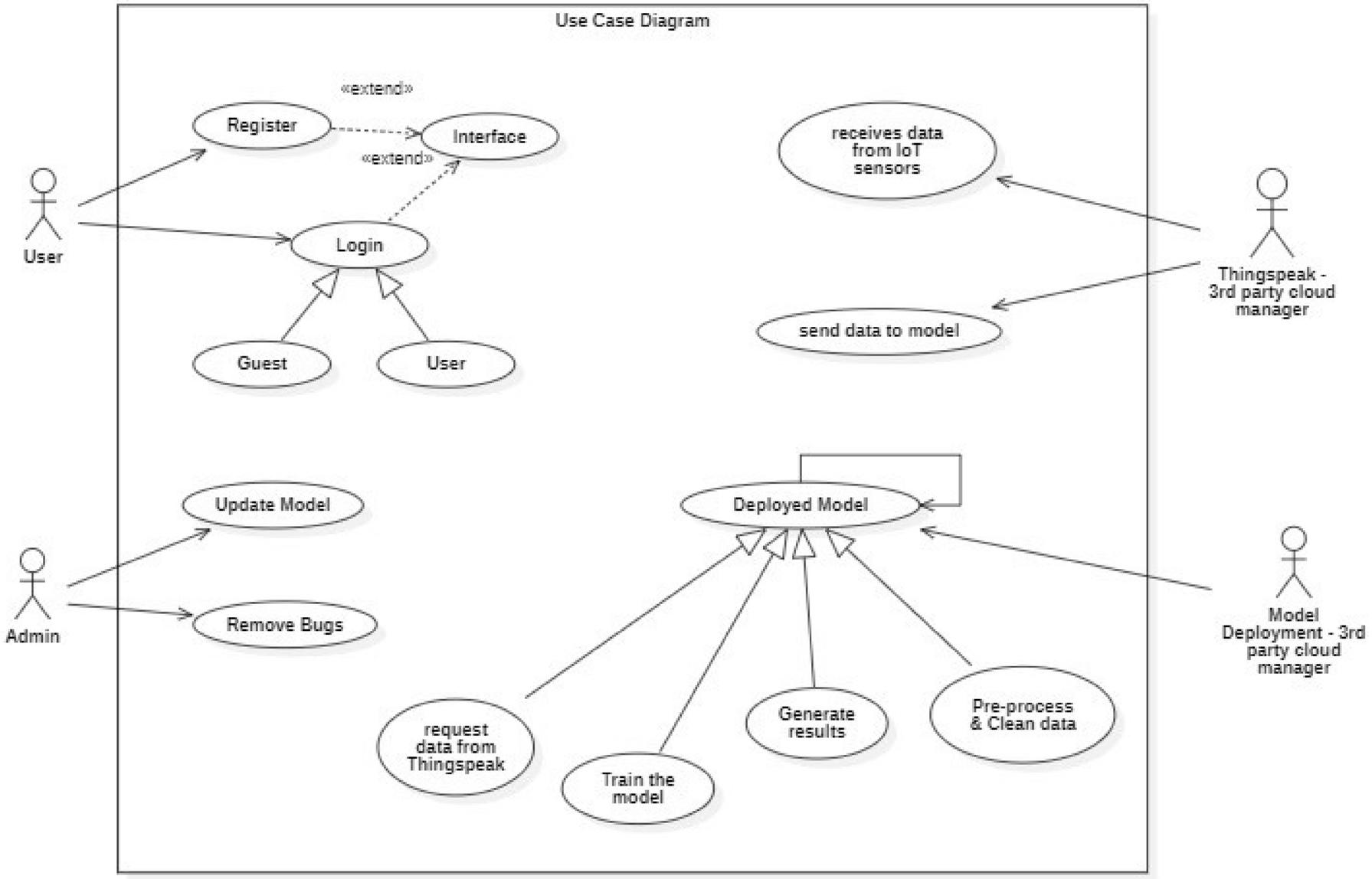
CIRCUIT DIAGRAM



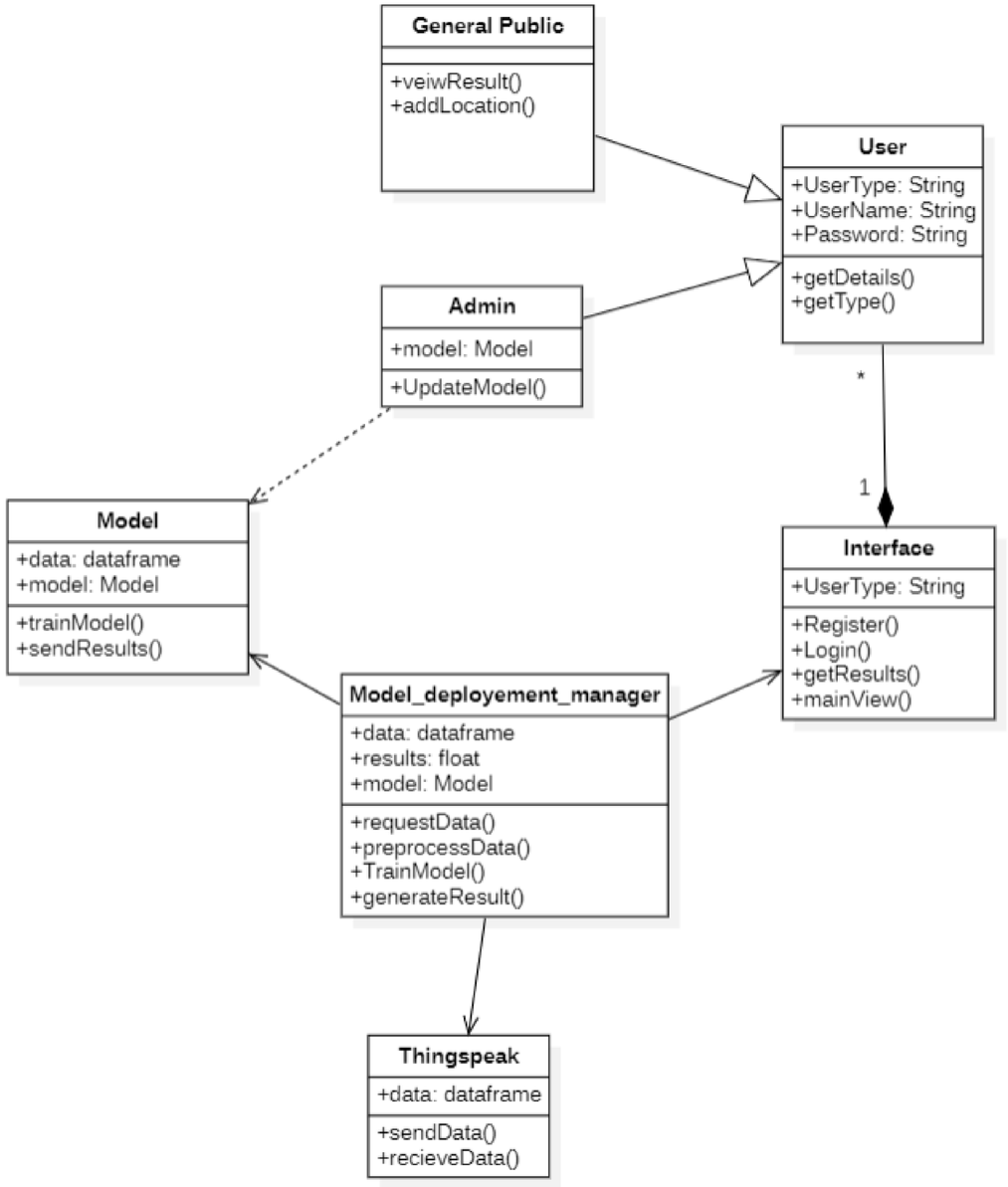
CIRCUIT DIAGRAM

- Using 2 models of ESP boards. ESP 32 and ESP 8266.
- DSM501A Dust Sensor, GP2Y1010AU0F and PMS5003.
- For the sake of the circuit diagram DSM501A is used.
- 3 components are 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.

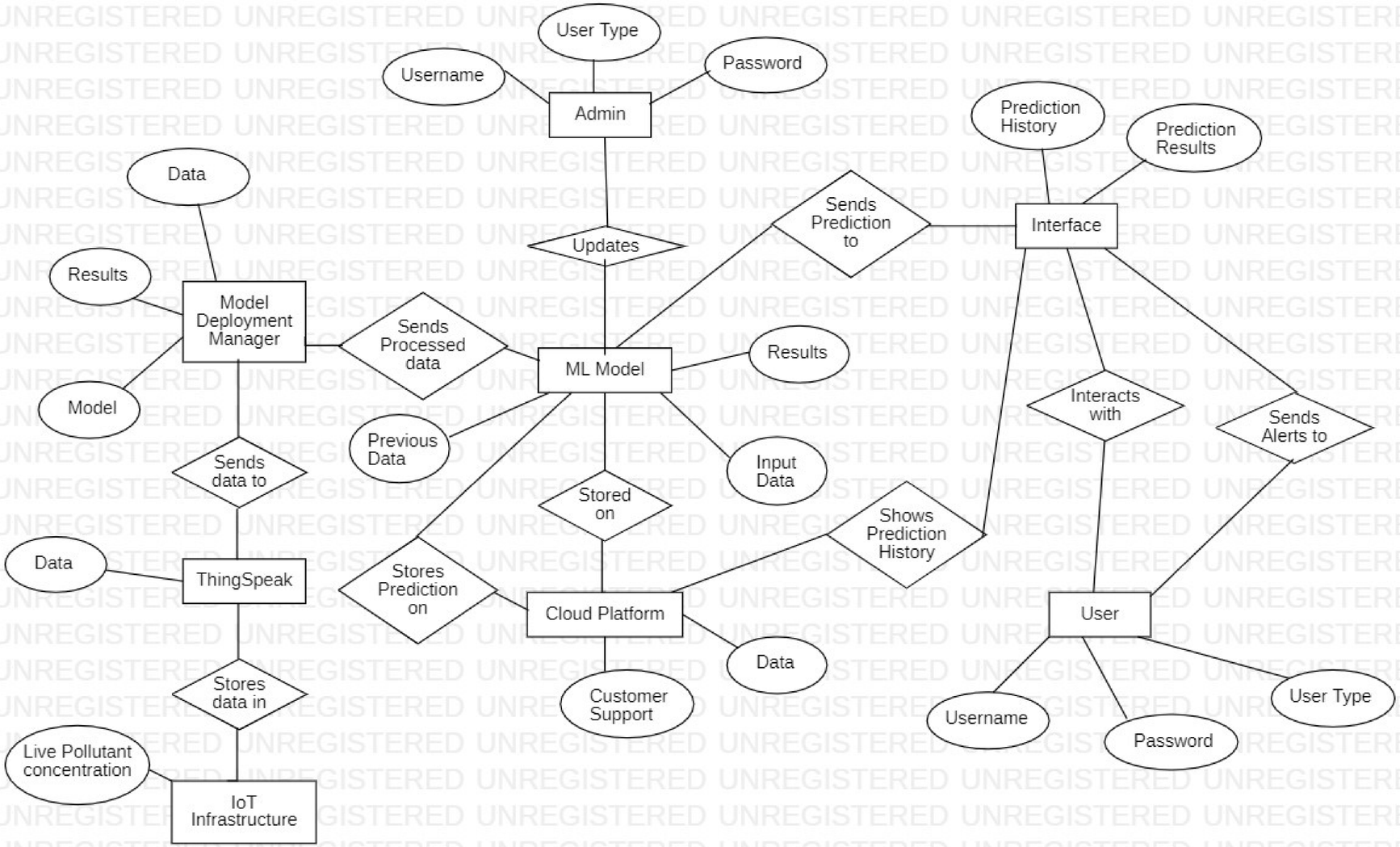
USE CASE DIAGRAM



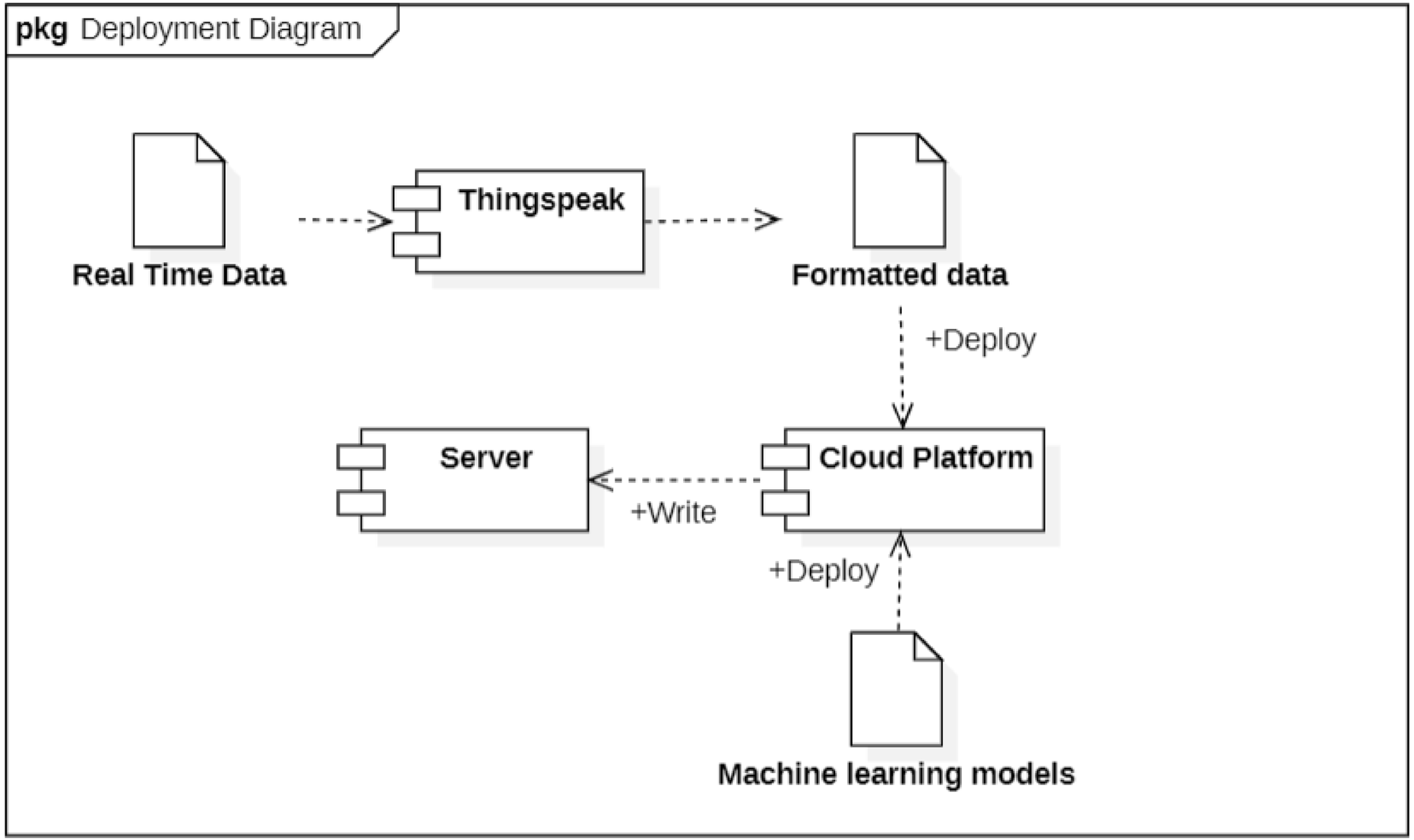
CLASS DIAGRAM



ER DIAGRAM


User Type
Username
Password
Admin
Updates
Data
Results
Model Deployment Manager
Model
Sends data to
Data
ThingSpeak
Stores data in
Live Pollutant concentration
IoT Infrastructure
Sends Processed data
ML Model
Previous Data
Stored on
Stores Prediction on
Cloud Platform
Customer Support
Data
Results
Input Data
Sends Prediction to
Shows Prediction History
Prediction History
Prediction Results
Interface
Interacts with
Sends Alerts to
User
Username
Password
User Type

DEPLOYMENT DIAGRAM



DATA PREPARATION & VISUALISATION

PROJECT PROGRESS

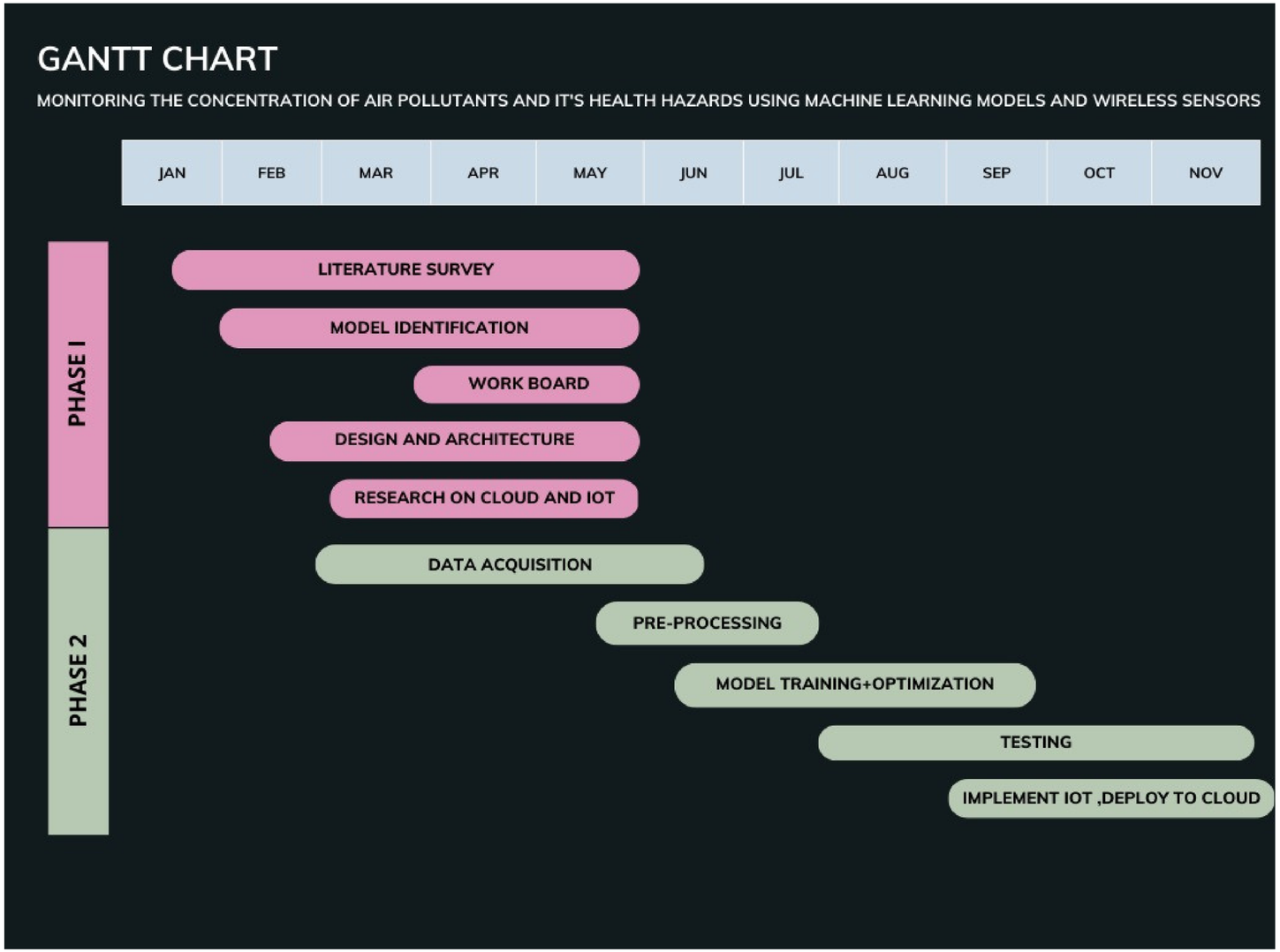
Completed

- Performed literature review.
- We have one large dataset ready for lung cancer.
- We have finalized the ML model.
  - bidirectional adaptive LSTM + ARIMA
- We have created architectural and structural diagrams.
- Chosen cloud platforms
  - ThingSpeak
  - Hugging face, Streamlit
- Exploratory Data Analysis

To- Do

- Find more datasets of countries similar to India
- Setup of the IoT infrastructure to collect the data
- Setup the cloud platforms
- Implementation of the ML model.

PROJECT TIMELINE



CONCLUSION

There has been considerable progress on the project in the Phase-I of Capstone.

The feasibility and practicability has been researched of achieving the project goals after referencing several research papers, acquired a few datasets of relevance including a large dataset for Lung Cancer and researched into standards for combining and extending datasets.

A few viable cloud platforms have also been looked at, ThingSpeak, Streamlit and HuggingFace to name a few, but not limited to the mentioned options.

The ML model has been finalised: bidirectional adaptive LSTM + ARIMA and the hardware

The sensor station/IoT infrastructure needs to be created; architectural and structural diagrams have already been created to visualize the proposed systems [Software and Hardware].

After thorough research, the team is confident in their approach and the direction they are heading in.

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