PROJECT REQUIREMENTS SPECIFICATION

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

UE20CS390A – 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

TABLE OF CONTENTS

1. Introduction

This document specifies the software requirements for a project that predicts the effects of monitored air quality on lung cancer.

This document is intended to

- Explain the purpose and features of the project
- The feasibility of the project
- How the ML model would respond to various air quality concentrations. The document gives a clear and concise description about the hardware and software interfaces, constraints and features that meet the user stories.

1.1 PRODUCT SCOPE

The product scope entails the following:

- Creation and optimization of an ML algorithm to predict the probability of a person contracting diseases like Lung cancer and Skin cancer.
- Measures the concentration of various pollutants in the air like PM10, SO2 and NO2 using wireless sensors.
- Uploading of the data collected by the wireless sensors to the cloud and pass it through the ML model

2. Literature Survey or Existing System

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

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.

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)

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 .

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

Approach & Results:

- They use an ensemble of ML models (ANN, Deep Learning, Decision Tree)
- This paper assesses the relationship between COVID-19-related deaths, economic growth, PM10, PM2.5, and NO2 concentrations in New York state.

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.

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

Approach & Results:

- They use ANN models for predicting the levels of air pollutants 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.

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.

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.

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

Approach and 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.

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.

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

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

Approach and Results:

The objective of this paper is to explore the association between household ventilation and lung cancer. Epidemiologic and household ventilation data were collected using a standardized questionnaire. Unconditional logistic regression was employed to estimate adjusted odds ratios (ORadj) and their 95% confidence intervals (CI).

Advantages:

- The data collected does consider one's family history on lung cancer, basic demographic factors, socioeconomic status, tobacco smoking history, alcohol consumption, 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.

Limitations:

- Selection bias and recall bias may exist in the study as the data is concentrated only on the local population.
- 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.

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

Approach & Results:

- They use and ensemble of models like Logistic Regression, Random Forest, SVM and Gradient Boosting and perform benchmarking on them.
- 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.

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

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.

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

Approach and Results:

- This article focuses on the correlation between PM and skin diseases, along with related immunological mechanisms.
- Increased PM levels are highly associated with the development of various skin diseases via the regulation of oxidative stress and inflammatory cytokines.
- Therefore, antioxidant and anti-inflammatory drugs may be useful for treating PMs induced skin disease

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.

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

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

Approach and Results:

- This paper analyzed the links between cancer mortality, socioeconomic 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

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.
- Explored the potential spatial association between socioeconomic and lifestyle factors.

Limitations:

- 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 different levels of air pollution.

3. Product Perspective

3.1. Product 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.

3.2. 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

3.3. 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 that you prefer can be used as long as it supports git or some other way to have a shared codebase.

- **Cloud**: We will use a few cloud platforms for this, which might evolve as time progresses and we find the pros and cons of each platform. Thingspeak will be used to save the data from the sensor stations and any cloud platform such as AWS Glue, Google Cloud Dataflow or Azure Data Factory to process our data and host our model.

3.4. General Constraints, Assumptions and Dependencies

- **Inaccurate Datasets**: We will not be able to verify that the datasets we're utilizing to train our model are authentic

- **Obtaining data**: We will need to ask regulating authorities for many of the datasets because they won't be available publicly

- **Scalability:** As the number of sensors and the amount of data collected increase, the ML model may become more complex and require more computational power

- **Data Complexity:** The data produced by IoT devices is often unstructured and provides a limited perspective along with the massive size

- **Energy Consumption:** The amount of energy used by IoT technology is significant and it needs to be running constantly.

- **Data privacy and security:** Air quality data can contain sensitive information about individuals or organizations, and it is essential to ensure that the data is protected from unauthorized access or misuse.

3.5. Risks

- **Data privacy and security**: The project involves collecting sensitive health data from individuals. There is a risk of data breaches and unauthorized access to this information, which could lead to privacy violations and compromise the integrity of the study.

- **Data accuracy and bias**: 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.

- **Ethical concerns**: There are ethical considerations related to collecting health data from individuals, especially if the data is being used for commercial purposes

- **Technical issues**: 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.

4. 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 user-friendly 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**: The system should be able to analyze the collected data and calculate the probability of lung cancer and other related diseases based on the concentration of pollutants in the air.

5. External Interface Requirements

5.1. User Interfaces

- 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 <decide interval> . It can send each measurement to the database, preferably immediately. The model can rerun with the new received data every <longer interval>..

- All error messages will be sent to the cloud platform. Some errors can be displayed on the LCD.

5.2. 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

5.3. Software Requirements

- **Data collection and preprocessing tools**: You will need tools to collect air pollutant data from various sources, such as air quality sensors. You 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**: We will use a cloud computing platform like AWS, GCP, ThingSpeak or Azure to store and process large volumes of data and to deploy our machine learning models.

- **Database management systems**: We will need a database management system like MySQL, PostgreSQL, or MongoDB to store and manage the collected data.

5.4. Communication Interfaces

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.

6. Non-Functional Requirement

- **Performance**: The system should be able to handle large amounts of data and provide real-time processing of data.

- **Maintainability**: The system should be easy to maintain and update, with minimal downtime.

- **Security**: The system should be secure and protect data privacy, as IoT devices are often vulnerable to attacks.

- **Scalability**: The system should be scalable and able to handle a growing number of devices and data points.

6.1. Performance Requirement

- The ML model should provide high accuracy with minimum error, approximately 95% accuracy with minimum error.
- **Scalability**: The system should be scalable and able to handle an increasing number of devices and data streams.
- **Maintainability**: The system should be easy to maintain and update, with minimal downtime.
- **Performance**: The system should be able to handle large amounts of data and provide real-time processing of data.
- The system should be compatible with other devices like smartphones, tablets.
- **Reliability**: The system should be able to operate continuously without any downtime. It should also be able to recover from failures or errors without affecting the accuracy or reliability of the results.
- **Interoperability**: 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.
- **Cost-effectiveness**: The system should be cost-effective to develop, deploy, and maintain, with a clear understanding of the project budget and ROI.

6.2. Safety Requirements

- **Electrical Safety**: We will run all the components at the appropriate and recommended voltages.

- **Compliance**: The product should comply with relevant safety and environmental regulations in the regions it is being used in.

- **Data Security**: The product should use only trusted softwares and protect user data and prevent unauthorized access.

6.3. Security Requirements

- The system should be secure and protect data privacy, as IoT devices are often vulnerable to attacks.

- The data obtained for training the model must be from a verified and trusted source.

- The system should store and manage collected securely on the cloud

7. Other Requirements

- **Scalability**: This is dependent on the cloud platform of choice.

- **Maintainability**: The database and compute maintenance will be handled by the cloud platforms. The stations must be maintained by the creators.

- **Portability**: The station can be used anywhere with a stable wifi connection and power supply.

- **Ease of setup & ease of use**: Making the stations will be dependent on the user's proficiency with arduino and its related systems. The cloud systems will be usable to anyone with a basic understanding of cloud systems.

Appendix A: Definitions, Acronyms and Abbreviations

- **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: 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/2071-1050/14/12/7015]

[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]

[9]] **Increased risk of emergency department presentations for bronchiolitis in infants exposed to air pollution**

By: Elisa Gallo, Silvia Bressan, Simoneta Baraldo
[https://onlinelibrary.wiley.com/doi/full/10.1111/risa.14007]

[10] **Deep learning architecture for air quality predictions**

By: Xiang Lee, Ling Peng, Yuan Hu [https://link.springer.com/article/10.1007/s11356-016-7812-9]

[11] **Air Quality Prediction using Machine Learning Algorithms – A Review**

By: Tanisha Madan, Shrddha Sagar, Deepali Virmani
[https://ieeexplore.ieee.org/document/9362912]

[12] **Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models**

By: Doreswamy, Harishkumar KS, Yogesh KM
[https://www.sciencedirect.com/science/article/pii/S1877050920312060]

[13] **Link between environmental air pollution and allergic asthma**

By: Quingling Zhang, Zhiming Qiu, Kian Fan Chung
[https://jtd.amegroups.com/article/view/3582/pdf]