

# UE20CS390A – Capstone Project Review #2

(Project Requirements Specification and Literature Survey)

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

**Project ID:** 102

**Project Guide:** Prof. Saritha R

<b>Project Team:</b>	Aditi Jain	PES2UG20CS021
	Aditya R Shenoy	PES2UG20CS025
	Ananya Adiga	PES2UG20CS043
	Anirudha Anekal	PES2UG20CS051



## Outline

---

- Abstract
- Motivation Scope of the Project
- Suggestions from Review – 1
- Functional and Non - Functional Requirements
- Literature Survey
- Capstone (Phase-I & Phase-II) Project Timeline
- Conclusion
- References

## Abstract

---

### **Problem Statement**

Monitoring The Concentration Of Air Pollutants &  
Its Health Hazards Using Machine Learning Models

# Motivation and Scope of the Project

---

## Prognosis using an ML model

We will create and optimize an ML algorithm to predict the probability of a person contracting various conditions or diseases like Bronchitis, asthma, acne and Dermatitis.

## Measure pollutant concentration using wireless sensors

We will measure the concentration of various pollutants in the air like PM10, SO2 and NO2 using wireless sensors.

## Environment or local atmosphere Analysis

Using our sensors we can generate AQI. This can be used to understand how polluted our surroundings are.

## Deploying the model on a cloud platform

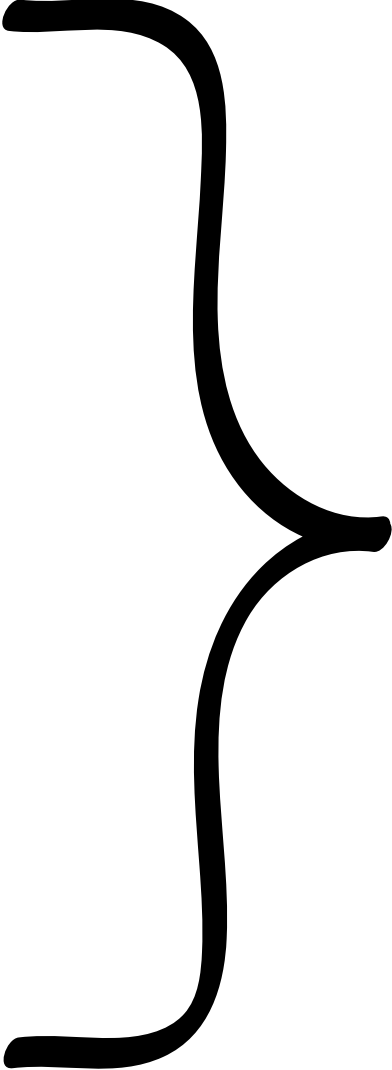
We will upload the data from the wireless sensors to the cloud and pass it through our ML model.

## Suggestions from Review – 1

---

### Suggestions

- Keep the focus more on prediction by machine learning models than IOT sensing
- Research further on the topic
- Limit the number of diseases



These suggestions have been kept in mind while further reading more papers and articles on the topic

# User Classes and Characteristics

---

## Smart home companies

Smart home businesses can incorporate it as an additional layer of detection for current devices

## Government

Governments can use this 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 can be informed about the levels of pollution and the risks associated with it to take preventive measures

# Constraints / Dependencies / Assumptions / Risks

---

## **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.

# Functional Requirements

---

The system should be able to collect real-time data on air pollutant concentrations from IoT sensors placed in different locations.

The system should send alerts if the concentration of pollutants exceeds a certain threshold.

The system should store and manage collected securely on the cloud

The system should be compatible with other devices like smartphones, tablets, etc.

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.



## Non - Functional Requirements

---

The ML model should provide high accuracy with minimum error

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

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.

# Literature Survey

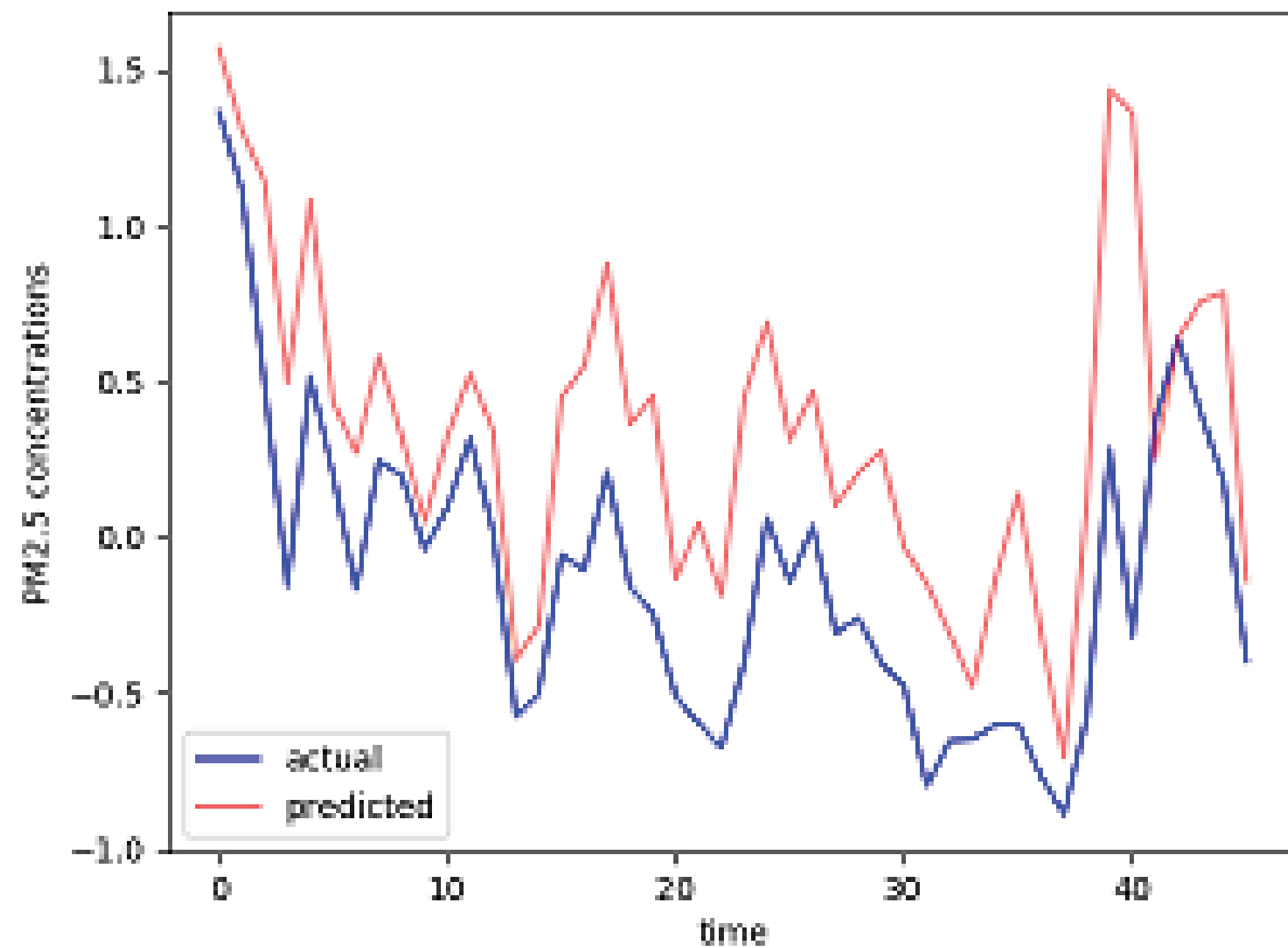
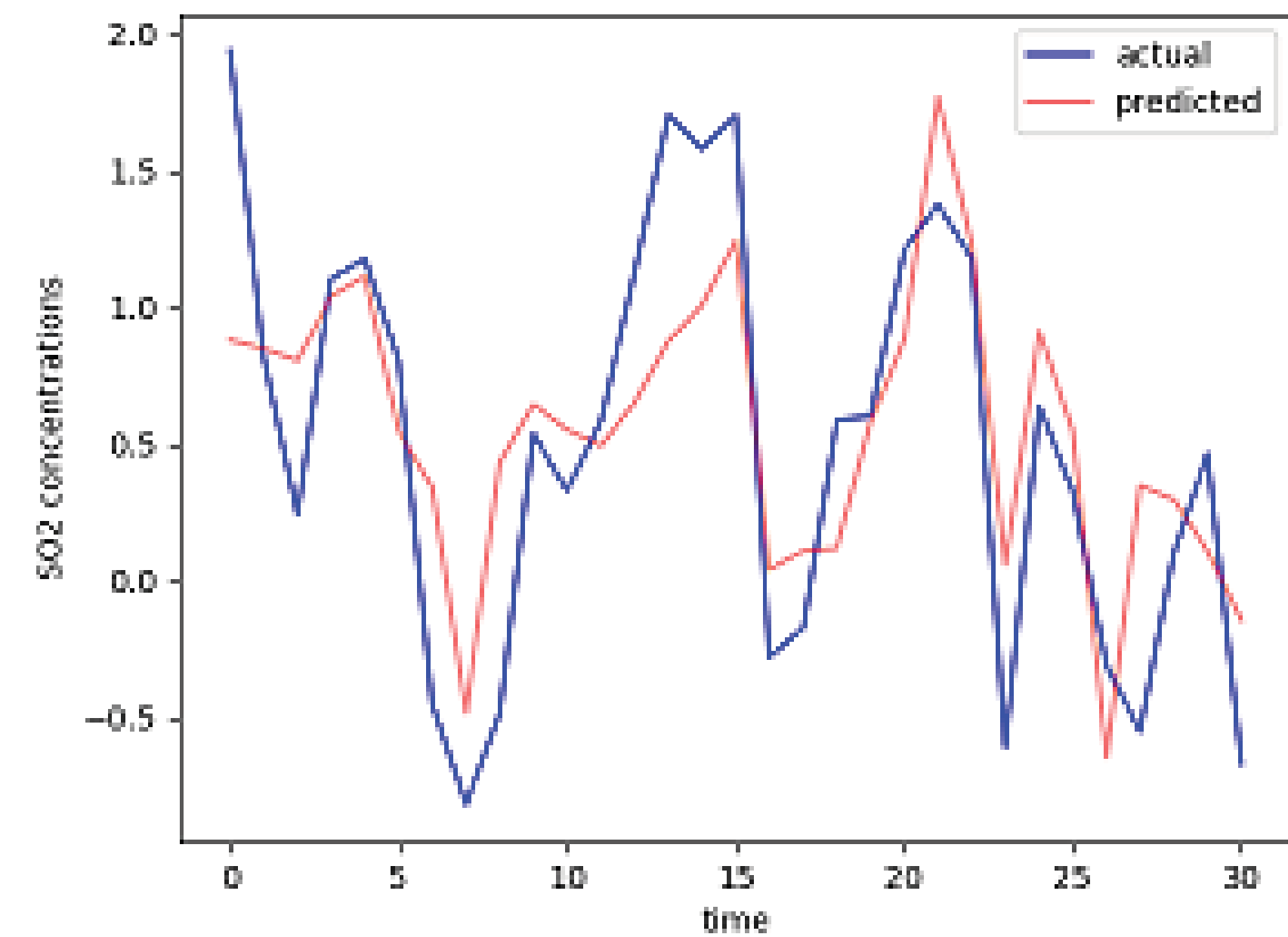
---

## Paper 1- Ananya Adiga

Paper Details	Objective of paper, Techniques/Methods	Advantages	Limitations
<p>Mohammad Marzouk and Mohammad Atef</p> <p><b>"Assessment of indoor air quality in academic buildings usng IOT and deep learnings":</b></p> <p>Mdpi(June 2022)</p>	<p>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.</p>	<ul style="list-style-type: none"> <li>• Uses real time data.</li> <li>• Finds the correlation between outdoor and indoor environment.</li> <li>• Concentrates on rural areas too, which might be affected by burning wood, construction particles and unpaved roads.</li> </ul>	<ul style="list-style-type: none"> <li>• Does not consider the family history with lung cancers</li> <li>• Does not consider the occupational exposures to carcinogens.</li> <li>• It is biased as the readings were collected only during summer when the humidity will be relatively high</li> <li>• The study was restricted to just a few primary pollutants and did not consider volatile organic compounds, NH3 and O3</li> </ul>

Paper Details	Objective of paper, Techniques/Methods	Advantages	Limitations
<p>Jin Z-Y, Wu M, Han R-Q, Zhang X-F, Wang X-S, et al. (2014)</p> <p><b>"Household Ventilation May Reduce Effects of Indoor Air Pollutants for Prevention of Lung Cancer: A Case-Control Study in a Chinese Population.</b></p> <p>PLoS ONE 9(7): e102685. doi:10.1371/journal.pone.0102685</p>	<p>The objective of this paper is to explore the association between household ventilation and lung cancer.</p> <p>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).</p>	<ul style="list-style-type: none"><li>• The data collected does consider one's family history on lung cancer, basic demographic factors, socio-economic status, tobacco smoking history, alcohol consumption, dietary history, and physical activity</li><li>• Active smokers, second hand smokers, carcinogens, tobacco smoking and high temperature cooking oil flames were considered for the study thereby covering a wide scope.</li></ul>	<ul style="list-style-type: none"><li>• Selection bias and recall bias may exist in the study as the data is concentrated only on the local population.</li><li>• The data about family history, age, gender is considered via a standardised questionnaire by interviewers, using quantitative data might have been effective instead.</li><li>• Have not considered occupational exposure</li><li>• The study is limited to IAQ and does not consider how outdoor pollutants are affecting IAQ.</li></ul>

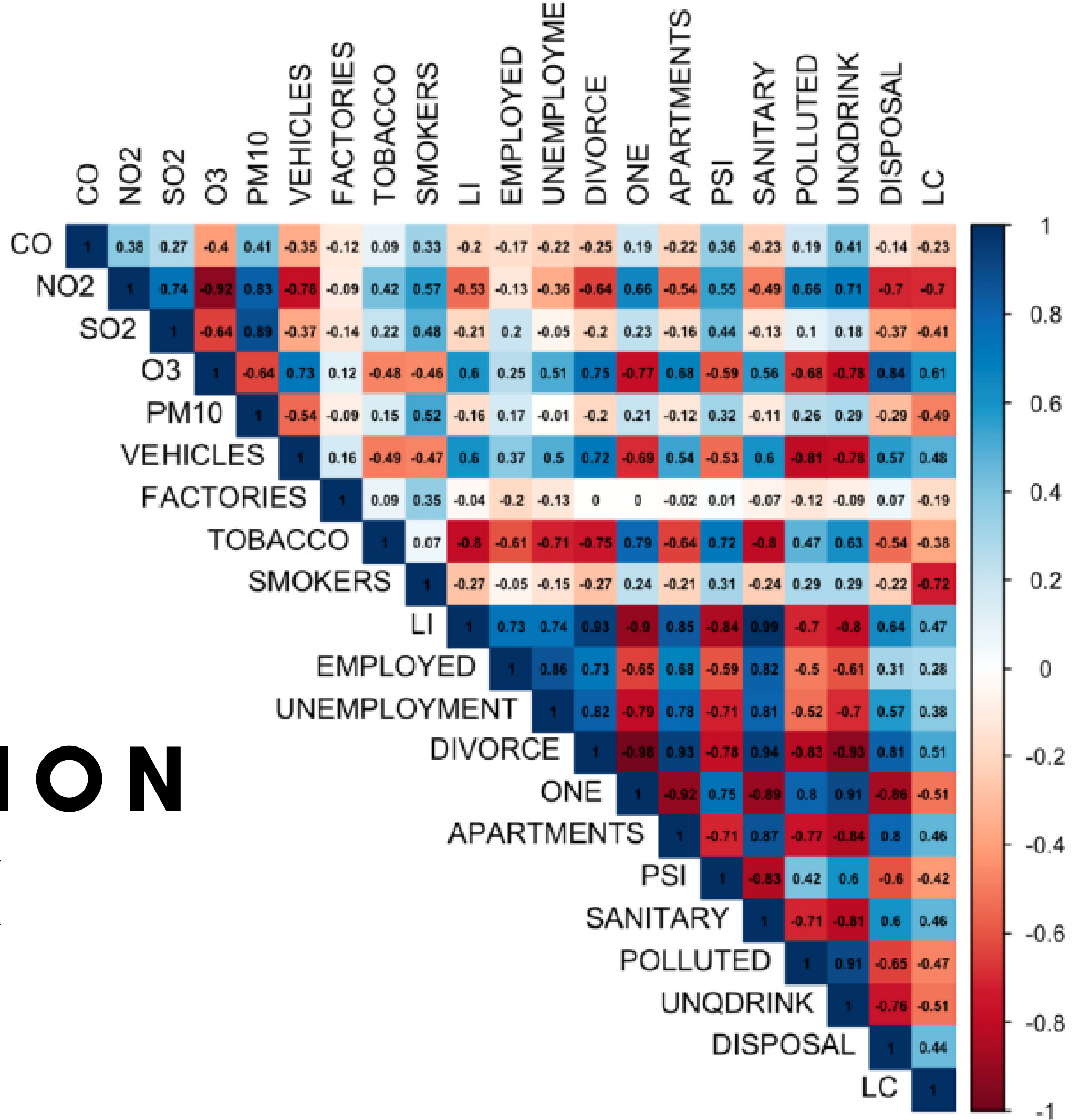
Paper Details	Objective of paper, Techniques/Methods	Advantages	Limitations
<p><b>An Application of IoT and Machine Learning to Air Pollution Monitoring in Smart Cities</b></p> <p>By: Muhammad Taha Jilani, Husna Gul A.Wahab Auckland University of Technology</p> <p>DOI: 10.1109/ICEEST48626.2019.8981707</p>	<p>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.</p> <p>They have proposed an IoT architecture that collects pollutants (PM10, SO2, NO2, O3, PM2.5) data and have made use of an ANN model for predicting the levels of air pollutants.</p>	<ul style="list-style-type: none"><li>• The paper takes into account the weather conditions as well. As air pollutants mixed with other factors like water/high winds cause differing effects</li><li>• The model they created has achieved a Root Mean Square Error of only 0.0128 for SO2 prediction and 0.0001 for PM2.5</li><li>• They have used two different methods (Pearson Correlation and ANN) for the weather correlation and prediction.</li></ul>	<ul style="list-style-type: none"><li>• Does not consider all the environmental weather conditions (Only wind speeds, temperature and humidity)</li><li>• The model only contains a single hidden layer and they have not run it only for 500 epochs</li><li>• The IoT infrastructure they have used is very crowded and complex.</li></ul>





Paper Details	Objective of paper, Techniques/Methods	Advantages	Limitations
<p><b>How Is the Lung Cancer Incidence Rate Associated with Environmental Risks? Machine-Learning-Based Modeling and Benchmarking</b></p> <p>By: Kung-Min Wang, Kun-Huang Chen, Shieh-Hsen Tseng</p> <p>National Taiwan University of Science and Technology</p> <p>DOI: 10.3390/ijerph19148445</p>	<p>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.</p> <p>The study aims to develop a predictive model to understand the impact of air pollutants on the disease.</p> <p>They make use of several ML models like Logistic Regression, Random Forest, SVM and Gradient Boosting</p>	<ul style="list-style-type: none"><li>• They perform benchmarking, i.e, comparing the performance of different ML models and provides insights on which the best one is .</li><li>• 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</li></ul>	<ul style="list-style-type: none"><li>• They have used a vast dataset and hence the completeness and accuracy of the data is a challenge.</li><li>• The paper doesn't consider several other factors like genetics and lifestyle factors that can also impact lung cancer incidence rates.</li><li>• They do not show causality. i.e, the study doesn't definitively show that air pollution causes lung cancer as other factors are not considered.</li></ul>

# CORRELATION MATRIX





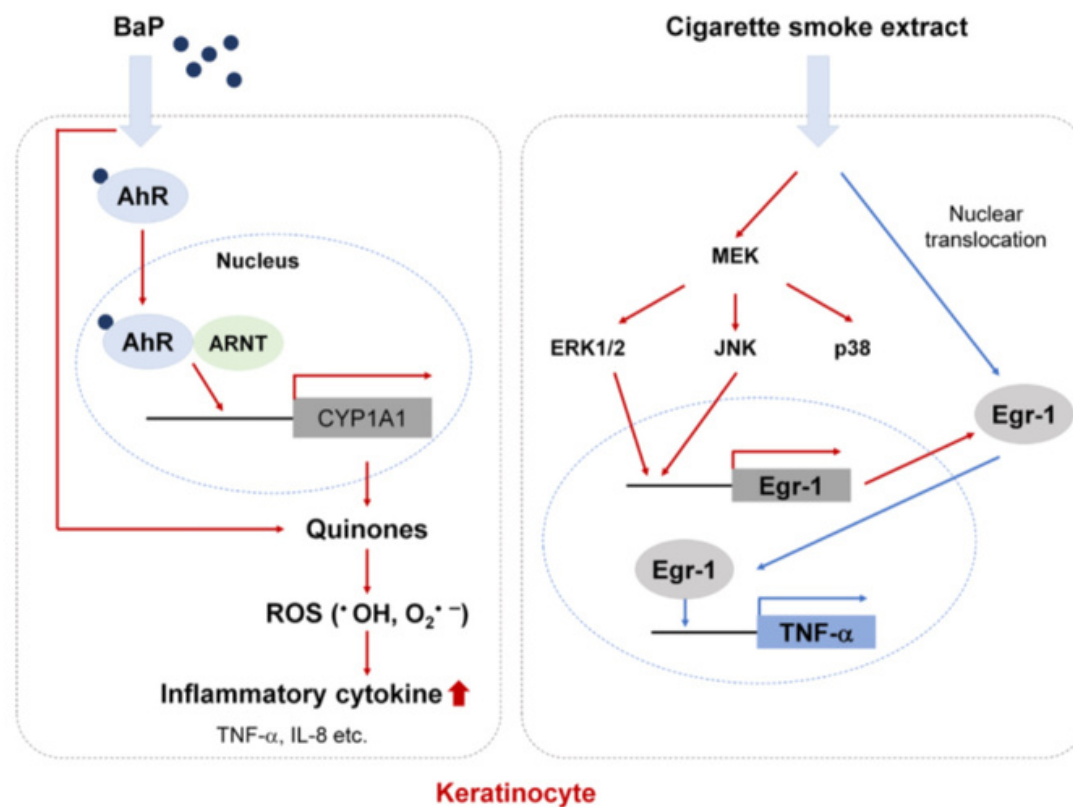
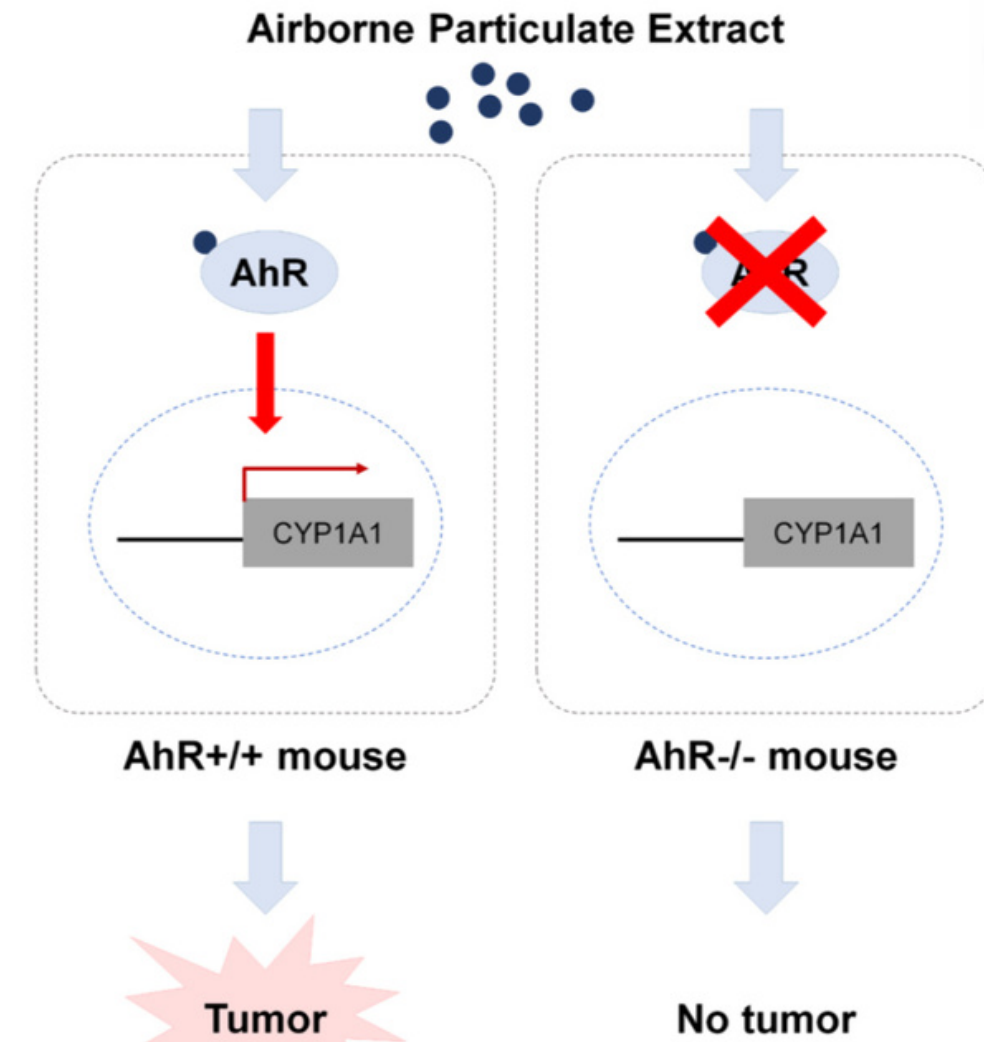
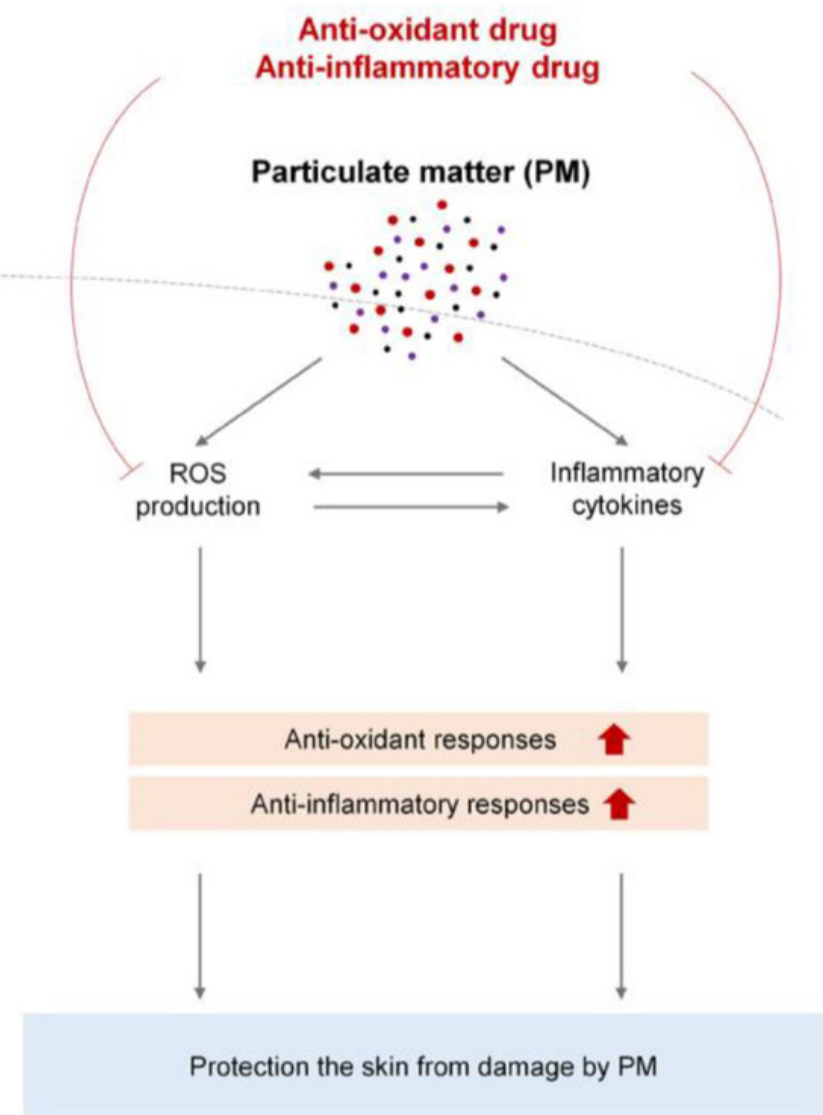
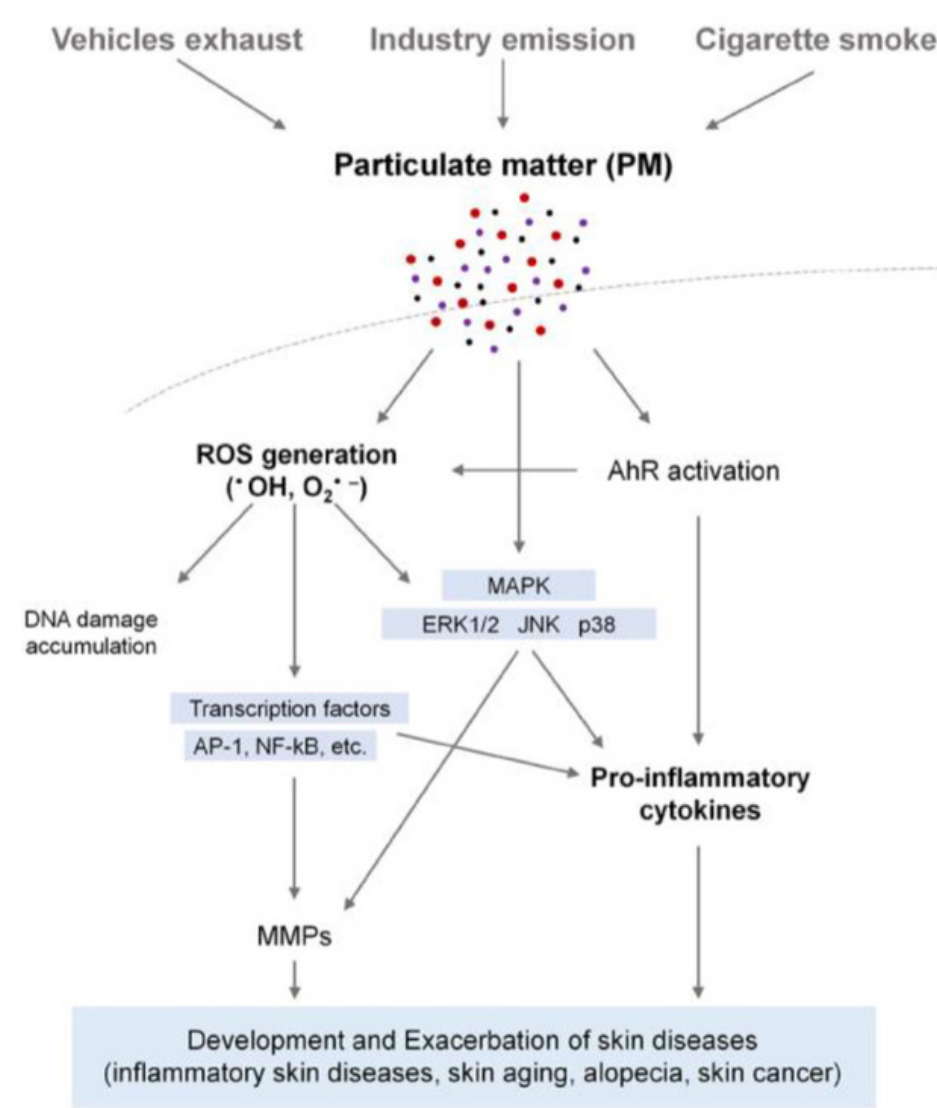
# Paper 5- Anirudha Anekal

Paper Details	Objective of paper, Techniques/Methods	Advantages	Limitations
<p><b>Determination of Air Quality Life Index (AQLI) in Medinipur City of West Bengal(India) During 2019 To 2020 : A contextual Study</b></p> <p>By: SAMIRAN RANA</p> <p>Dept. of Physiology, Research Scholar, Shri J.j.t University, Jhunjhunu, Rajasthan</p> <p>Doi: <a href="http://dx.doi.org/10.12944/CWE.17.1.12">http://dx.doi.org/10.12944/CWE.17.1.12</a></p>	<p>The objective of the paper is to show a provable reduction in life expectancy and the correlation to various air pollutants</p> <p>They do not use any ML model. They use the predefined method provided by various organizations like WHO, INAAQS etc.</p>	<ul style="list-style-type: none"><li>• They use air pollutants rather than AQI</li><li>• Speaks about the direct effects of PM on various parts of the body.</li><li>• Uses an expensive but very accurate, and lab tested/ approved sensor for PM 2.5</li></ul>	<ul style="list-style-type: none"><li>• Gathered data only from densely populated and heavily polluted areas. This prevents the analysis being relevant to people outside of these areas.</li><li>• Uses and external app for data gathering. Algorithm used is proprietary and not revealed (I think trade secret)</li><li>• Does not discuss the methods used. Just speaks about them like its common knowledge.</li></ul>

# Paper 6- Anirudha Anekal

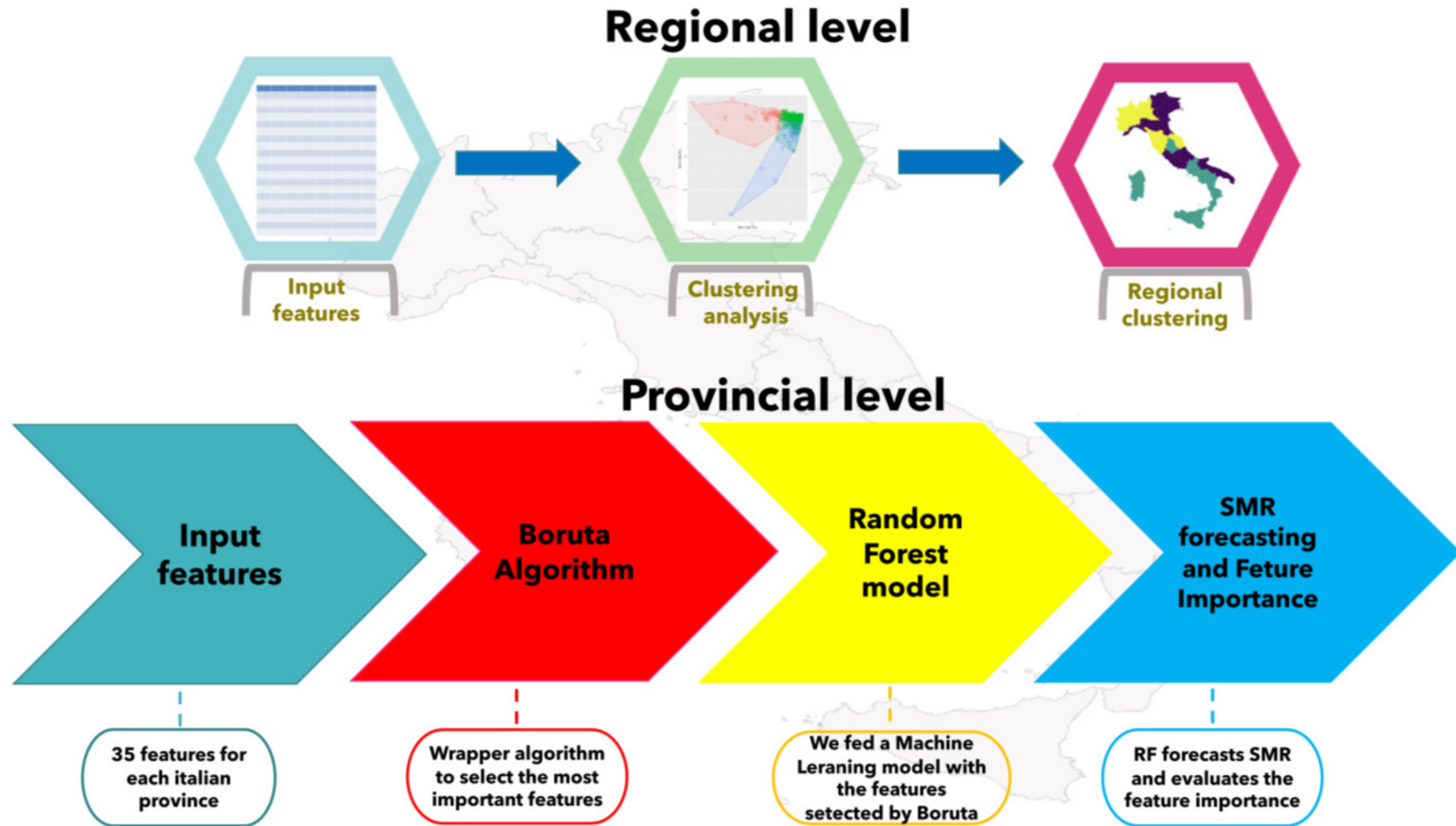
Paper Details	Objective of paper, Techniques/Methods	Advantages	Limitations
<p><b>The aim of this paper is to assess the relationship between COVID-19-related deaths, economic growth, PM10, PM2.5, and NO2 concentrations in New York state using daily city data through two Machine Learning experiments.</b></p> <p>By: Cosimo Magazzino , Marco Mele , Samuel Asumadu Sarkodie</p> <p>Department of Political Sciences, Roma Tre University, Italy Department of Political Sciences, University of Teramo, Italy Nord University Business School, Norway</p> <p><a href="https://doi.org/10.1016/j.jenvman.2021.112241">https://doi.org/10.1016/j.jenvman.2021.112241</a></p>	<p>This paper aims to prove a link between various air pollutants (like PM or NO2) and death.</p> <p>Considering the paper is from a department of business they also try to show a link to economic growth.</p> <p>Further it shows a link between unsustainable growth and air pollutants.</p>	<ul style="list-style-type: none"><li>• Considers individual pollutants.</li><li>• Very well thought out data gathering and cleaning.</li><li>• Techniques/Models used: ANN, Deep learning - Oryx Protocol, "Multiple regression" &lt;= derived from another paper, a specialized decision tree.</li><li>• The above are well suited for the type of predictions and the type of data we are using.</li></ul>	<ul style="list-style-type: none"><li>• It has been done mainly to covid related deaths.</li><li>• This has many inherent issues especially considering how little we truly know about covid.</li><li>• 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.</li><li>• Does not consider other underlying issues.</li></ul>

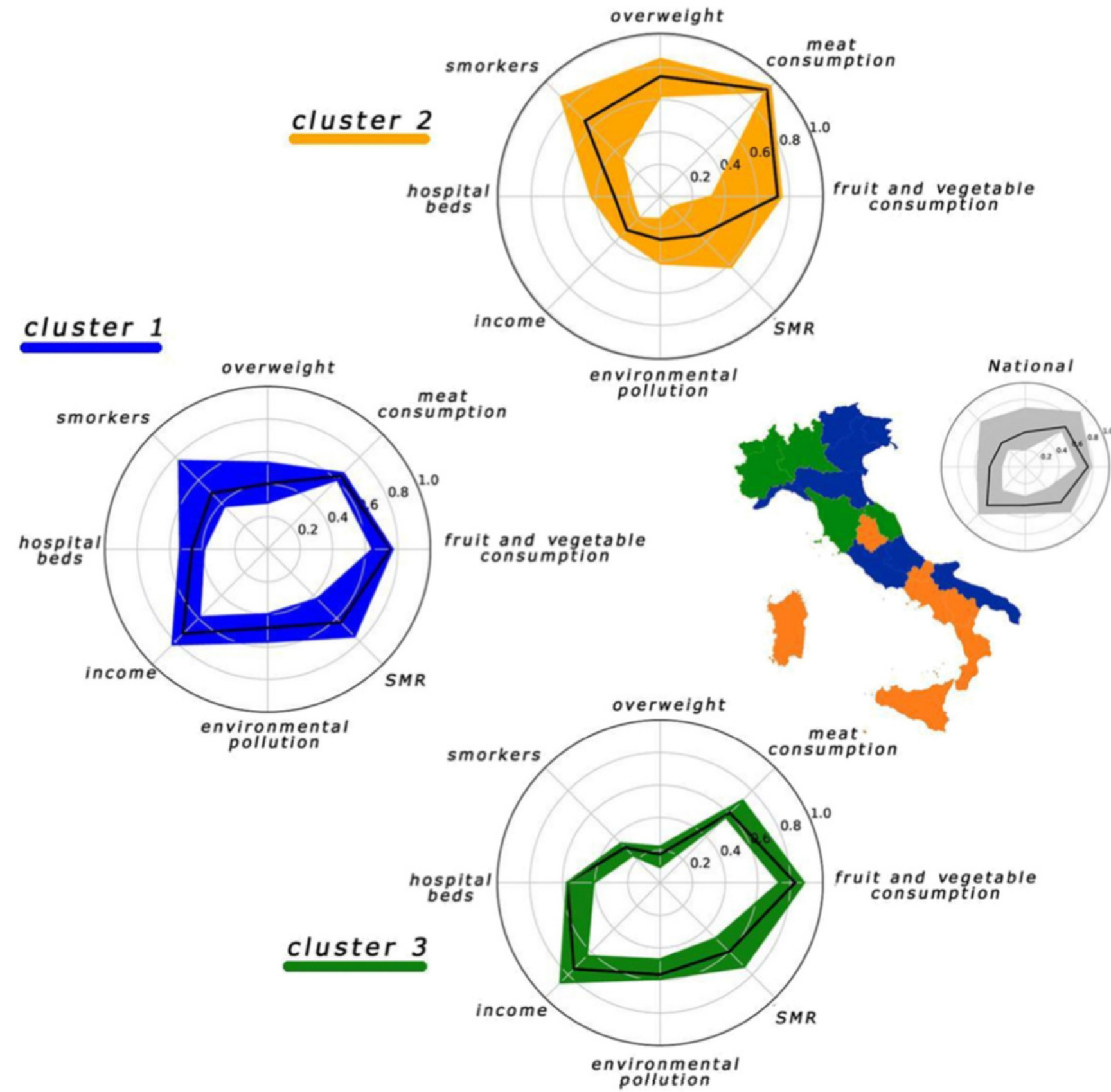
Paper Details	Objective of paper, Techniques/Methods	Advantages	Limitations
<p>Kim Kyung Eun, Cho Daeho, Park Hyun Jeong,</p> <p><b>Air pollution and skin diseases: Adverse effects of airborne particulate matter on various skin diseases,</b></p> <p>Life Sciences (2016),</p> <p>doi: 10.1016/j.lfs.2016.03.039</p>	<p>This article focuses on the correlation between PM and skin diseases, along with related immunological mechanisms.</p> <p>Increased PM levels are highly associated with the development of various skin diseases via the regulation of oxidative stress and inflammatory cytokines.</p> <p>Therefore, anti-oxidant and anti-inflammatory drugs may be useful for treating PM-induced skin diseases</p>	<ul style="list-style-type: none"><li>• This paper has listed in detail which air pollutants are known to cause which disease.</li><li>• It lists various diseases like Atopic dermatitis, Acne, Psoriasis, and skin cancer.</li><li>• It also discusses the conditions of skin aging, alopecia, and oxidative stress.</li></ul>	<ul style="list-style-type: none"><li>• The amount of statistics is less that is uses to prove its point.</li><li>• It doesn't discuss the correlation between these diseases, and how they affect the probability of another happening.</li></ul>





Paper Details	Objective of paper, Techniques/Methods	Advantages	Limitations
<p>Roberto Cazzolla Gatti, Arianna Di Paola, Alfonso Monaco, Alena Velichevskaya, Nicola Amoroso, Roberto Bellotti,</p> <p><b>The spatial association between environmental pollution and long-term cancer mortality in Italy,</b></p> <p>Science of The Total Environment,</p> <p><a href="https://doi.org/10.1016/j.scitotenv.2022.158439">https://doi.org/10.1016/j.scitotenv.2022.158439</a></p>	<p>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.</p> <p>Random Forest (RF) regression coupled with a Boruta feature importance analysis, K means clustering</p> <p>SMR forecasting and Feature importance, Regional cluster analysis</p>	<ul style="list-style-type: none"> <li>• It has taken into consideration all the different types of body parts that can get affected due to air pollution in extensive detail.</li> <li>• The data sources and the algorithms used have been discussed.</li> <li>• Explored the potential spatial association between socioeconomic and lifestyle factors</li> </ul>	<ul style="list-style-type: none"> <li>• Some punctual sources of pollution do not show a relation with any specific cancer type</li> <li>• 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.</li> </ul>





- A lot of the papers have considered similar machine learning models proving they are most efficient for usage in such a study.
- There are only a few papers listing in detail which type of pollutant causes which type of disease.
- Most papers either discuss only skin conditions or lung conditions, they don't discuss both.
- Most papers make use of AQI for their models.
- There is hardly any study on the interrelation of probability of diseases.



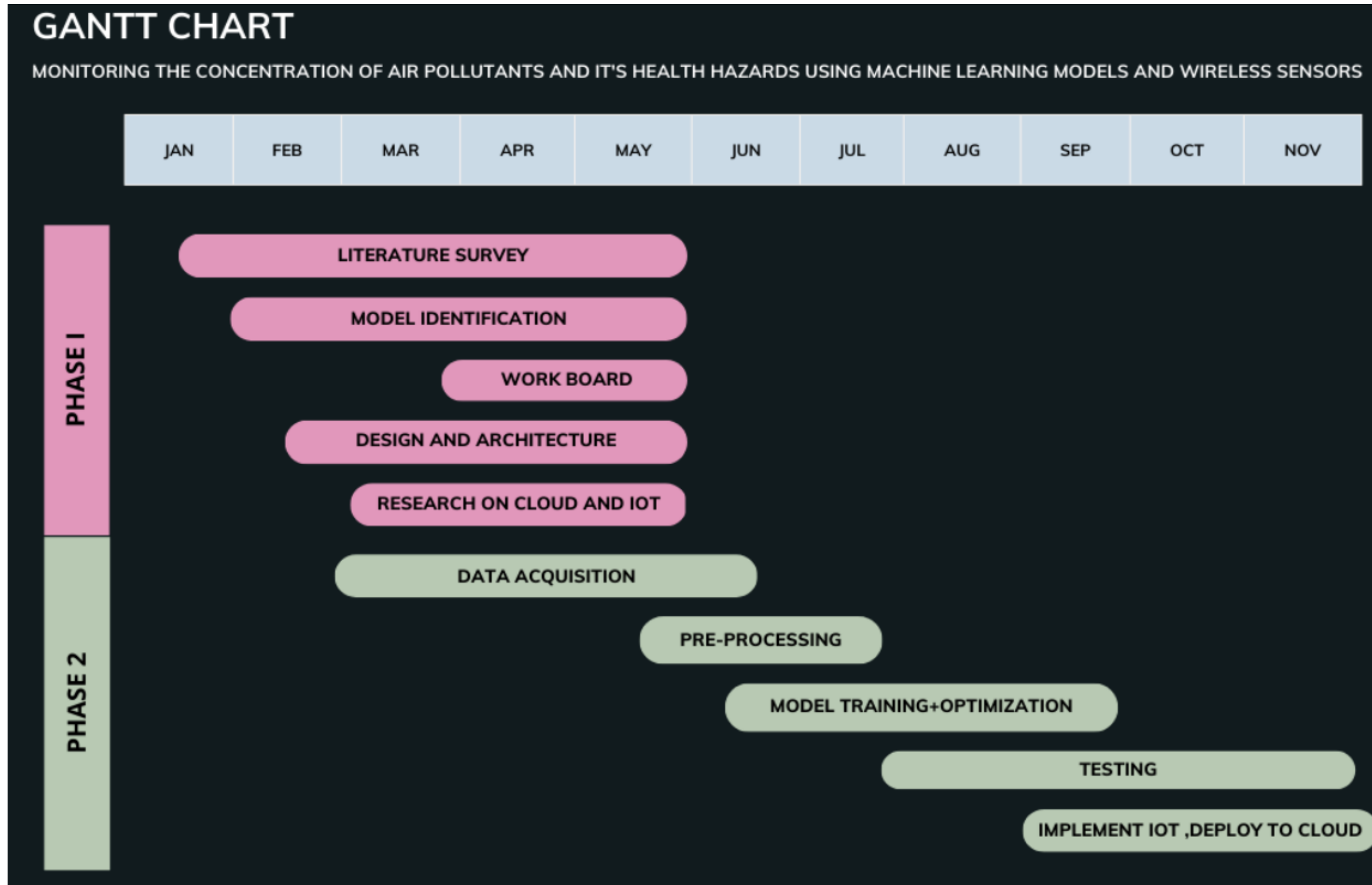
## Summary of Literature Survey

---

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 predictions based on dated and static data are infrequent.

# 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

[\[https://ieeexplore.ieee.org/document/8981707\]](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

[\[https://www.mdpi.com/1660-4601/19/14/8445\]](https://www.mdpi.com/1660-4601/19/14/8445)

[3] Assessment of indoor air quality in academic buildings using IOT and deep learnings

By: Mohammad Marzouk and Mohammad Atef

[\[https://www.mdpi.com/1667822\]](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](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/>]

## References

---

[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

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

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

Thank You