**Custom Dataset for Disease Prediction Based on Real-time Air Quality Data**

**Abstract**

This paper presents a dataset that was collected from 1000 individuals' disease records based on the air quality in each area. A specially designed air quality monitoring system was used to gather gas concentrations, convert them to ppb form for easier detection, and also calculate the air quality index using the presented gas data. Data from physical surveys and sensor values are included in the dataset. The MQ135 gas sensor, which is responsible for detecting harmful gases in the air and is integrated into an Arduino microcontroller, is used in the construction of the final product. For the gas sensors, data samples were taken at a rate of 9600 baud per action. The air quality index and gas concentration are gathered by the gas sensor, which shows the user's location's air quality and presented gases. Using this dataset, we can predict diseases caused by polluted air and try to reduce pollution to save lives.

* **Specifications Table**

**Table 1: Specification Table**

|  |  |
| --- | --- |
| **Subject** | Computer Science in Air Pollution |
| **Speciﬁc subject area** | Disease Prediction |
| **Data format** | Raw data of MQ135 gas sensor and people’s disease |
| **Type of data** | Table, each table having eighteen columns, CO2 Concentration (ppb), Ammonia (NH3) Concentration(ppb), Benzene Concentration(ppb) , NO2 Concentration(ppb), CO Concentration(ppb), Air Quality Index (AQI), Air Quality, Gender, Lung cancer, pneumonia, Stroke, Heart disease, Bronchitis, Chronic obstructive pulmonary disease, Asthma, genetically carried. |
| **Data collection** | Data was collected using a custom-built air quality monitoring system. We used MQ135 gas sensor, which gives CO2 Concentration (ppb), Ammonia (NH3) Concentration(ppb), Benzene Concentration(ppb), NO2 Concentration(ppb), CO Concentration(ppb), Air Quality Index (AQI), Air Quality data. The sensor is interfaced to the Microcontroller via the Arduino using a cardboard. |
| **Data source location** | Hamdard University Bangladesh, Jamaldi, Comilla, Voberchar,  Baluakandi, Daudkandi, Sonargao. |
| **Data accessibility** | Direct URL to data: https://github.com/codeMAHDI/air-quality-and-disease-prediction-data |

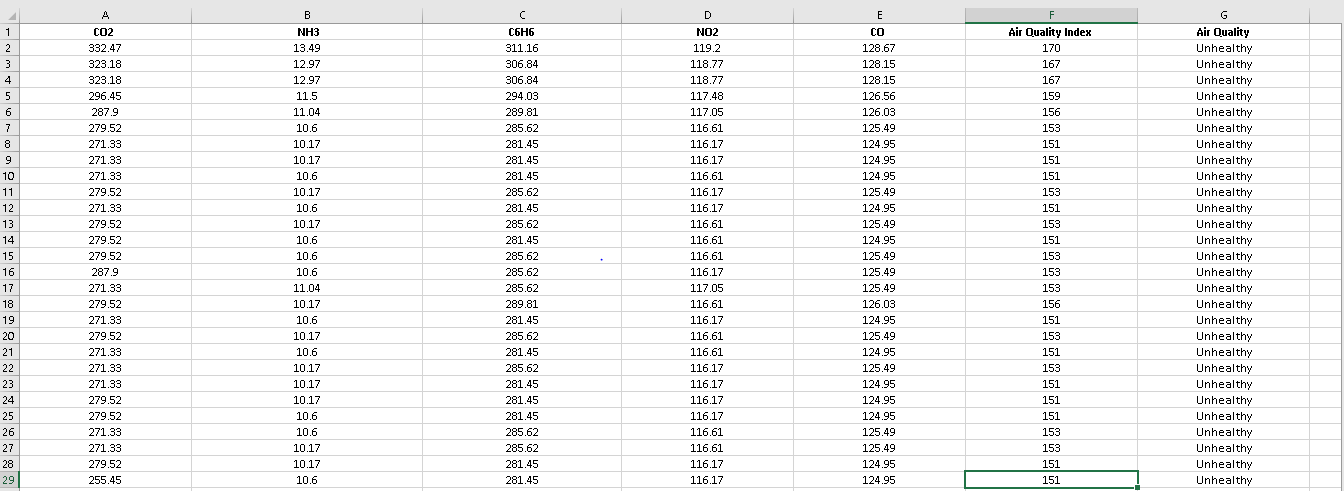
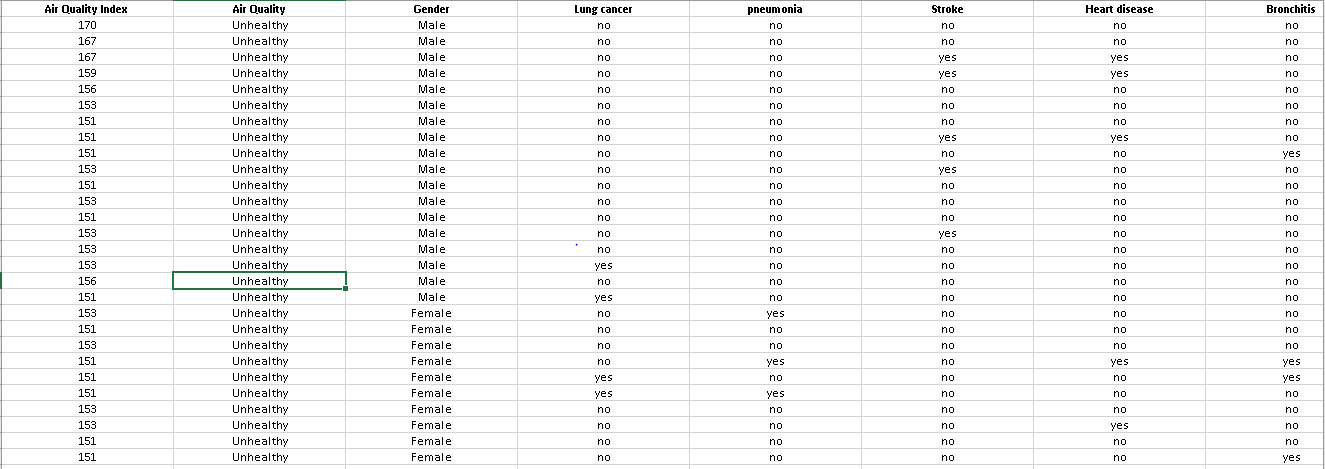
**1. Value of the Data**

* This is the only dataset available that uses real time sensor data, people’s survey data and has a total of 223 instances.
* There are no other gas concentration-based disease prediction datasets available, so the sensors are used to collect harmful particles of air and the disease data was collected from the physical survey.
* We have thoroughly analyzed the performance on various ML algorithms and obtained accuracies and precision in predicting diseases as high as 86%.
* Highly scalable as the device used to collect the data is a low-cost device that can be easily used.

**2. Data Description**

* **Format of the csv ﬁle**

Fig.1 gives the snapshot of the excel ﬁle. Column A, B, C, D, E, gives the CO2, NH3, C6H6, NO2, CO gas concentrations. Which were recorded by the sensor. The next two column i.e. F and G give Air Quality Index and Air Quality values and H gives the gender values. Last 8 columns i.e. I, J, K, L, M, N, P gives the different disease data and O gives the genetic information if there was a connection or not.



**Fig. 1. Sample dataset**

**3. Experimental Design, Materials and Methods**

In this dataset, the data was collected from a total of 1000 people, some of whom are listed in Table 1. The collection of data from the gas sensors was done using an Arduino code. The people were asked about their affected diseases by the air pollution. The air quality monitoring device on his or her side while asking all the questions. The circuit placement of the sensor is shown in Fig. 2 below. This system is built around a gas sensor that is used to detect all gases present in the air. The sensor code is developed specifically for IoT applications. We used a MQ135 gas sensor, an Arduino microcontroller that gives the needed gas concentrations, and software named CoolTerm for data collection. The sensor is interfaced to the microcontroller using a card board. The collected data was stored in a csv file.

* **Survey Statistics**

In this survey, data was collected physically. We go to the industrial areas and talk about air pollution with the local people, ask them a few questions, and record their statements for further purposes. The questionnaires were listed in Table 2. We differentiated the percentage of people affected by the disease between males and females. And the Table 3 is for Genetical history. Fig.3 shows the survey statistics.

* **Questionnaires and details**

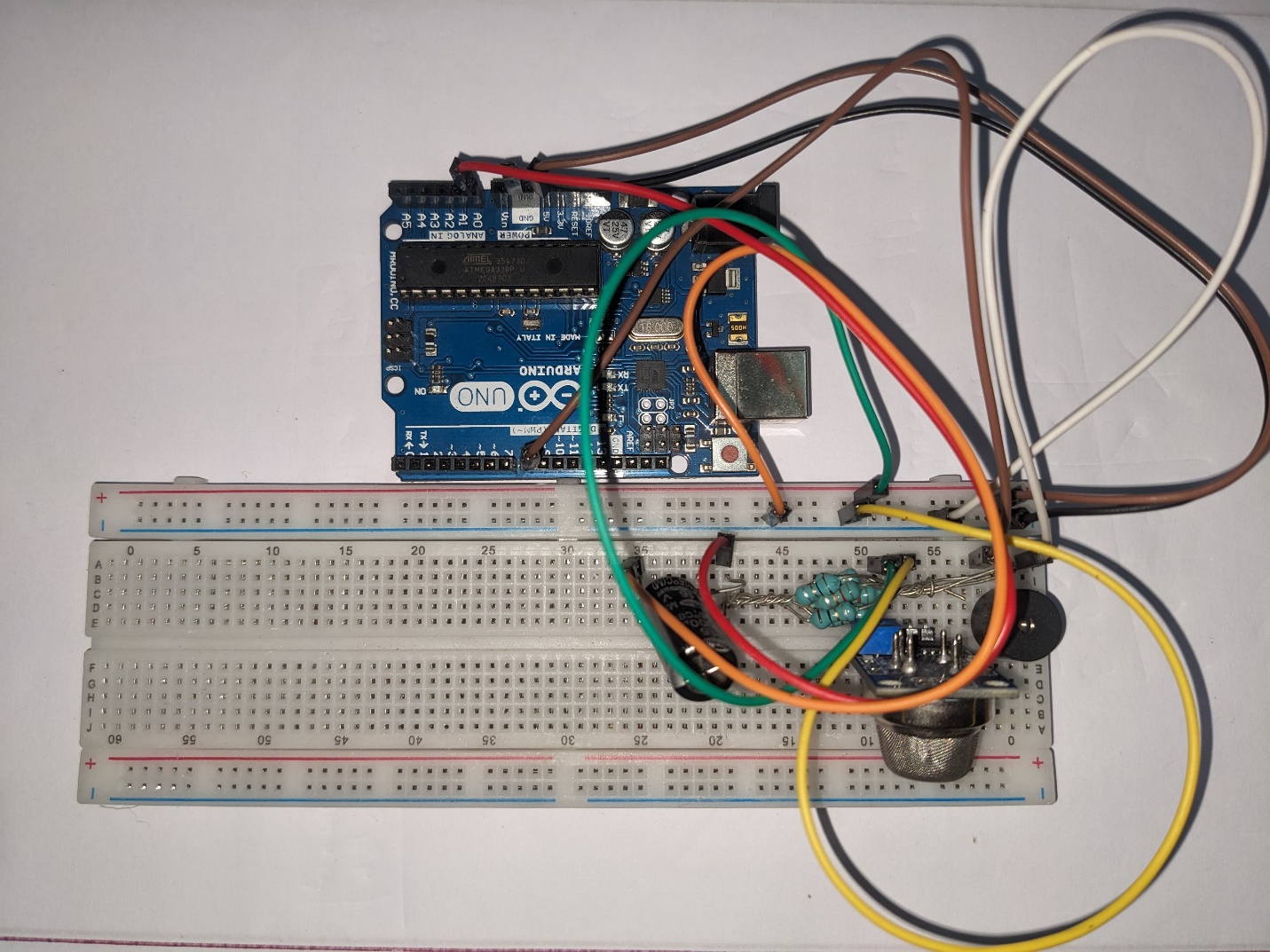
**Table 2: Questionnaires and details**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl.** | **Questionnaires** | **Male** | **Female** |
| **01** | Is the hot weather effecting you? | 60% | 40% |
| **02** | Have you experienced discomfort or headaches during the recent heat wave? | 55% | 55% |
| **03** | Have you ever suffered suffocated for the polluted of air? | 62% | 38% |
| **04** | Do you have lung cancer or have you had it previously? | 27.3% | 12.3% |
| **05** | Do you have or have had pneumonia? | 3.5% | 21.3% |
| **06** | Have you ever experienced a stroke? | 18.5% | 1.8% |
| **07** | Do you have heart disease or have you had it before? | 13.0% | 20.1% |
| **08** | Do you have or have had bronchitis? | 10.6% | 19.8% |
| **09** | Do you have chronic obstructive pulmonary disease or have you had it before? | 0.2% | 5.5% |
| **10** | Do you have asthma, or have you had it before? | 26.9% | 19.3% |

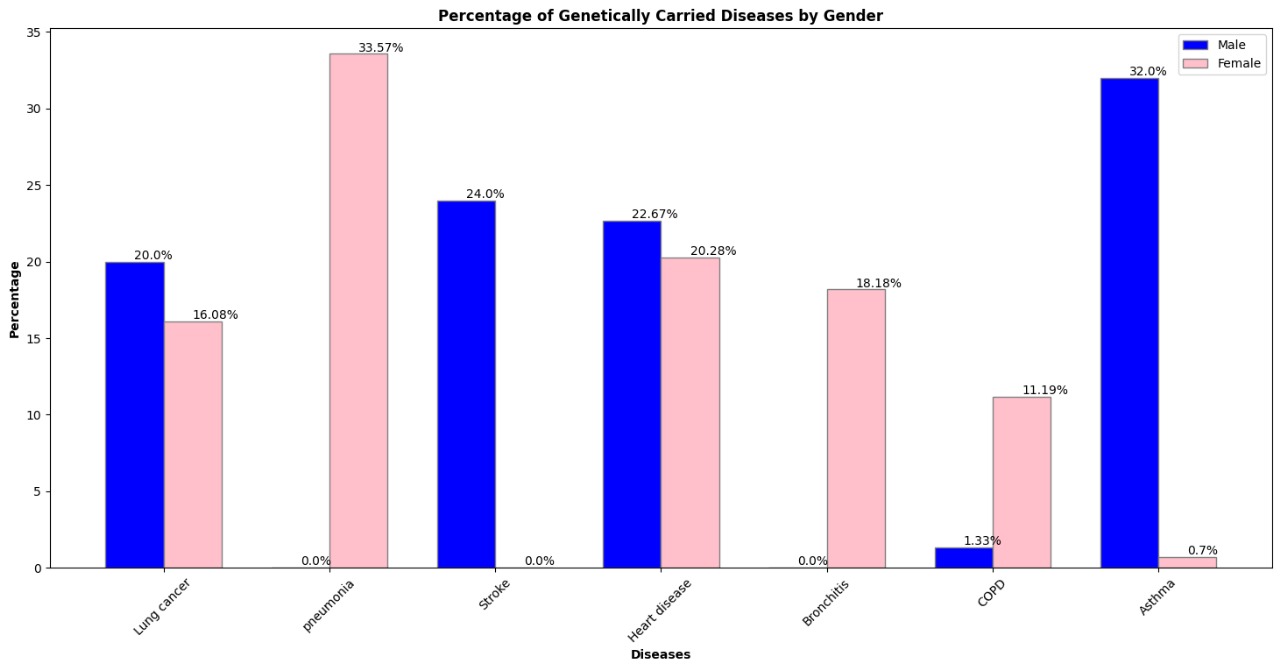
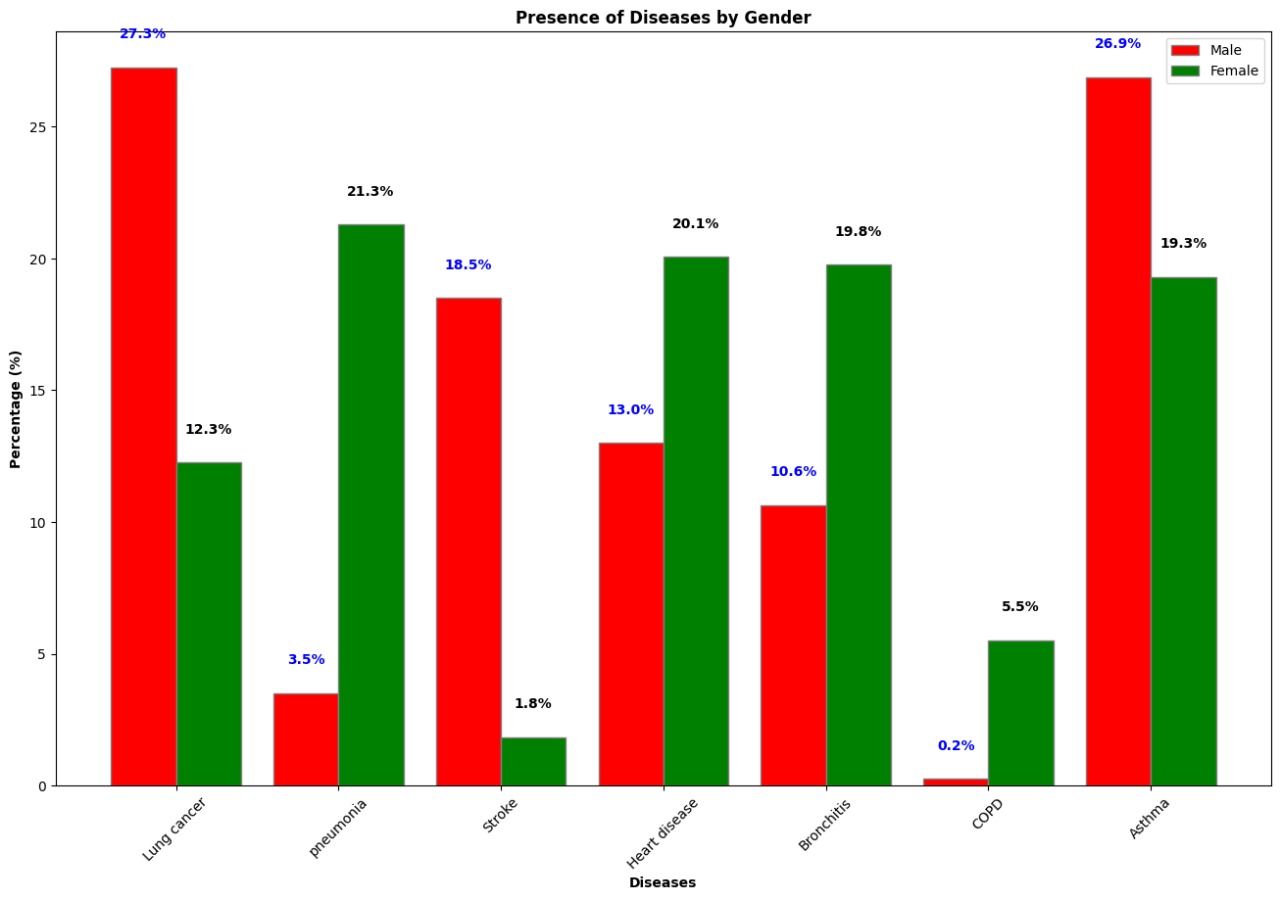
* **Family History**

**Table 3: Genetical History**

|  |  |  |
| --- | --- | --- |
| **Diseases** | **Male** | **Female** |
| Lung Cancer | 20% | 16% |
| Pneumonia | 0% | 33.57% |
| Stroke | 24% | 0.0% |
| Heart Disease | 22.67% | 20.28% |
| Bronchitis | 0.0% | 18.18% |
| Chronic Obstructive Pulmonary disease | 1.33% | 11.19% |
| Asthma | 32% | 0.7% |



**Fig. 2. Device used for collecting data and its internal circuit.**

**Fig. 3. Survey Statistics**

* **Survey details**

**Table 4: survey details**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Gender** | **Lung cancer** | **pneumonia** | **Stroke** | **Heart**  **disease** | **Bronchitis** | **Chronic obstructive pulmonary disease** | **Asthma** | **Genetically**  **carried** |
| Male | no | no | no | no | no | no | yes | no |
| Male | no | no | no | no | no | no | no | no |
| Female | yes | no | no | no | no | no | no | yes |
| Female | no | no | yes | yes | no | no | no | no |
| Female | no | no | no | yes | no | no | no | yes |
| Male | no | no | yes | yes | no | no | no | no |
| Male | no | no | yes | yes | no | no | no | no |
| Male | no | no | no | no | no | no | no | no |
| Male | no | no | no | no | no | no | yes | yes |
| Female | no | yes | no | no | no | no | no | no |
| Female | no | no | no | no | no | no | no | no |
| Male | no | no | no | no | no | no | yes | yes |
| Male | no | no | yes | yes | no | no | no | yes |
| Male | no | no | no | no | yes | no | no | no |
| Male | no | no | yes | no | no | no | no | no |
| Female | no | no | no | no | no | no | no | no |
| Female | no | yes | no | yes | yes | no | no | no |
| Male | no | no | no | no | no | no | no | no |
| Male | no | no | no | no | no | no | no | no |
| Female | yes | no | no | no | yes | no | yes | no |
| Female | yes | yes | no | no | no | no | no | no |
| Male | no | no | no | no | no | no | no | no |
| Male | no | no | yes | no | no | no | yes | no |
| Male | no | no | no | no | no | no | no | no |
| Male | yes | no | no | no | no | no | no | no |
| Female | no | no | no | no | no | no | no | no |
| Male | no | no | no | no | no | no | yes | no |
| Male | yes | no | no | no | no | no | yes | no |
| Female | no | no | no | yes | no | no | no | no |
| Female | no | no | no | no | no | no | yes | no |
| Female | no | no | no | no | yes | no | yes | no |

**Limitations**

Not applicable.

**Ethics Statement**

As our work involved on human subjects, we have monitored the AQI through our system and collected disease data by physical survey where the statement was recorded and two witness were presented. And later this data has been used in machine learning for prediction.

**Data Availability**

**GitHub link: https://github.com/codeMAHDI/air-quality-and-disease-prediction-data**

**Credit Author Statement**

**Moshiur Rahman**

**Mumtahina Mehjabin**

**Acknowledgements**

No specific grant was given for this research by public, private, or nonprofit funding organizations.

**Declaration of Competing Interest**

The authors declare that none of the work reported in this study could have been influenced by any known competing financial interests or personal relationships.

**References**

**[1**] D. Verma, S. Dhul, R. Saini, R. B. Dubey "IoT Based Air Pollution Monitoring System”, International Journal of Innovative Research in Engineering & Management (IJIREM), Volume 5, Issue 3, May-2018

**[2]** K. Kumar, B. P. Pande “Air pollution prediction with machine learning: a case study of Indian cities”, International Journal of Environmental Science and Technology), Volume 20, Issue 15, May 2022.

**[3]** Perumal Saravanan Pillai “Solar Smart Air Pollution Monitoring Framework Using Arduino and IoT Applications”, Volume 6, Issue 04 April- 2024.

**[4]** Vijayakumar Sajjan, Pramod Sharma "Research on an Iot Based Air Pollution Monitoring System”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume 8, Issue 9S2, July 2019.

**[5]** Anand Jayakumar, Praviss Yesyand, Venkstesh Prashanth, Ramkumar, “IoT Based Air Pollution Monitoring System”, International Research Journal of Engineering and Technology (IRJET). Volume 08, Issue 03, Mar -2021.