IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Prediction Of Indoor Air Quality Using A Statistical Web Tool

[1]Sabari Kesav S, [2]Balakrishnan C , [3]Mahalakshmi J , [4] Vinay M , [5] Gobi Ramasamy

[1]Research Intern, Department of Computer Science, Christ University, Bengaluru, Karnataka, India;

[2]Associate Professor, Department of Computer Science, Christ University, Bengaluru, Karnataka, India;

[3]Assistant Professor, Department of Computer Science, Christ University, Bengaluru, Karnataka, India;

[4]Associate Professor&Head, Department of Computer Science, Christ University, Bengaluru, Karnataka, India;

[5]Associate Professor, Department of Computer Science, Christ University, Bengaluru, Karnataka, India;

Abstract: It is increasingly being recognized that indoor air quality has an impact on human health. The increasingly serious indoor air pollution caused by PM, VC, and CO in recent years has attracted plenty of attention, which the world had not known before because of the politics, industry, and network caused by the urbanization development and industrialization.

A statistical method to predict IAQ has been developed in this work, and an easy-to-use website-based framework, including the statistical model, was implemented. The system can provide air quality predictions from one to seven days in advance by integrating statistical forecasting methods with a stable IAQI calculation process.

A flexible pipeline for data processing that supports different CSV-based inputs constitutes the basis of the approach. It is based on statistical forecasting with residual correction, scaling of pollutants, and calculation of the IAQI according to the Central Pollution Control Board (CPCB). The predictions are subsequently served to the users via a modern Flask application, enabling dataset uploads and visualisation of both numerical and categorical predictions. The system is an economical instrument that transforms statistical modeling information into a helpful decision-making framework for a policymaker with an emphasis on accessibility, reproducibility, and application.

Index Terms - IAQI, website-based framework, Flask application, statistical forecasting, and indoor air quality.

1. INTRODUCTION

The quality of indoor air and human health have a direct link, which often remains undermined. According to the WHO (World Health Organization), exposure to poor air quality contributes significantly to respiratory disorders, cardiovascular complications, and reduced life expectancy [1]. While considerable attention has been devoted to outdoor pollution, studies show that people spend nearly 90% of their time indoors [2]; as a result, the prediction and monitoring of IAQ has emerged as a critical research area.

Indoor environments often contain a mixture of harmful pollutants. Fine particulate matter (PM2.5) that is often invisible to the naked eye can infiltrate so deeply into the lungs that it may enter the bloodstream and cause long-term respiratory illness and cardiovascular risks[1]. Meanwhile, larger particles such as PM10 exacerbate asthma and bronchitis. Carbon monoxide (CO), a gas with no colour

or order, combines with hemoglobin in blood, which causes a depletion in the amount of oxygen the blood carries to vital organs, causing huge concerns [1]. Temperature and humidity, which are classified as environmental factors and are not pollutants, do influence microbial growth and pollutant behaviour, which indirectly affect indoor air quality [4]. Conventional IAQ monitoring is limited by stationary sensors that only provide a snapshot of pollutant levels at a specific time. These basic detection systems, which simply provide readings of the current conditions, are not intelligent; rather than providing homes and workplaces with preventative actions, they leave them in a reactive mode. Preventing future consequences through prediction is handled by statistical and ML models [3]. Such a capability is essential to update filteration systems, discuss ventilation strategies, and alert vulnerable populations of health hazards beforehand to minimize damage.

The goal of such a tool is to create a robust and user-friendly model for IAQ prediction through statistical modeling. It is designed as part of an IAQI (Indoor Air Quality Index) calculation method that integrates with a Flask-based web tool. This makes it accessible to non-technical users while still delivering reliable predictions. Individuals and organizations can obtain future forecasts with minimal effort by inputting IAQ data and taking future precautions to safeguard health.

2. LITERATURE SURVEY

The quality of indoor air has recently received much-needed attention as a result of the increasing percentage of employment residing indoors, since exposure to pollutants in these environments has surpassed outdoor levels [2]. Studies expanding over time have developed a link with particulate matter to increased mortality, cardiopulmonary disease, and lung cancer [10,11,15,16], while short-term spikes worsen asthma and bronchitis cases [14,17,18]. Ozone exposure has also been associated with elevated mortality risk [13].(*Figure 2.1*)

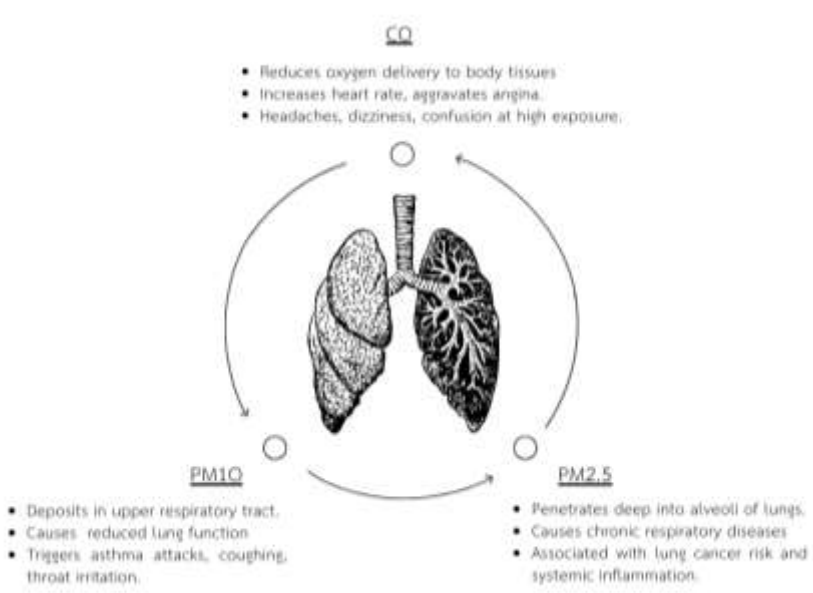

Figure 2.1 : Pollutants and Health Effects

Sensor monitoring has traditionally been relied on to comprehend indoor air quality, but these systems often give present data readings rather than precautions or measures for improvement [4] *(Figure 2.2)*. However, constant data capturing has become feasible due to IoT innovations; these systems commonly offer only reactive reporting [9]. Studies done on source apportionment reveal the need for incorporating weather conditions for predictions, taking into account the influence climate clustering and pollutant composition have on IAQI variability [6,12].

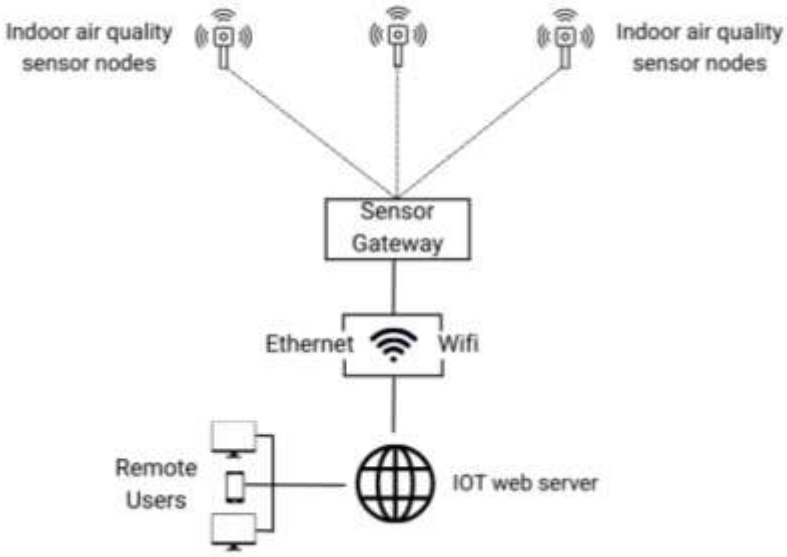

Figure 2.2: IoT Sensor Working Diagram

There are three main streams of methodological advancements in IAQ prediction. Firstly, regression and ARIMA, such statistical models, have been proven reliable in providing an interpretable baseline in cases of data scarcity [7,19]. Second, machine learning approaches, including random forests, decision trees, and SVMs, offer stronger nonlinear modeling and generally outperform simple regressions [5,19]. Third, deep learning models like RNNs and LSTMs have been proven to produce light short-term accuracy with the availability of sufficient data, but these systems limit widespread deployment due to their computational needs.[7]. Hybrid pipelines, that merge breakdown approaches with deep models like BiLSTM or attention systems to enhance robustness, have emerged as a remedy to nonlinear dynamics and data noise [3,8].

Communication of IAQ predictions is another active area of research. Interactive web platforms and mapping tools allow forecasts to be translated into actionable insights for both the public and policymakers [9].

Inference

The literature highlights several requirements for effective IAQ prediction: pollutants control to avert health hazards[1,10–18], integration of meteorological context to improve robustness [6,7,12,19], and balanced modeling strategies that trade off predictive accuracy with computational efficiency [3,5,7,8,19]. Effective presentation layers like web display of clearly labelled dashboards to ensure the predictions are transformed into meaningful insights [4,9,20]. Collectively, the reviewed literature indicates that even though established methodologies exist, most cases are either limited to digitalized reading or remain too constrained to incorporate in practice. This gap underscores the need for lightweight, standards-aware forecasting frameworks such as the one proposed in this study, which combines robust pollutant processing, a compact prediction module, and deployable web integration.

3. EXISTING SYSTEM

The current methods in indoor air quality monitoring are mostly descriptive. PM2.5, PM10, CO, and VOCs online [4]. Although beneficial, these models offer only a snapshot and do not predict future states. Efforts incorporating predictive modeling have included Random Forest, as well as regression-based approaches with mixed results [5].

Moreover, these systems are generally inflexible since they need standardized datasets in fixed formats up to a single byte of individual data and the entirety of the data. For databases, real-world data are plagued with missing values and have inconsistent encodings or different labels for features, which makes these models not useful. Added to that is that predictive algorithms, developed in research, hardly ever even become available for the broader public.

Thus, there is an immediate need to create the next generation of IAQ systems that would be predictive, adaptive, and applicable in the field.

4. MATERIALS AND METHODS

3.1 Dataset

The dataset contains recorded Air Quality Data of the Shillong Region of Northeast India. The year-long data spans the full cycle of seasonal variations, which is critical for understanding and forecasting air quality trends. The dataset follows the Central Pollution Control Board (CPCB) standards for IAQI values. The major pollutants recorded in the dataset are CO, PM2.5, and PM10, along with temperature and relative humidity serving as climate aspects.

Datetime	NO2	Ozone	CO	SO2	AT	RH	PM2.5	PM10
01-07-2022 00:00	2.1755232	1.171428632	233.0435842	4.965645867	23.33662516	72.36183352	11.88380751	19.94398114
01-07-2022 01:00	2.1755232	1.171428632	233.0435842	4.965645867	23.33662516	72.36183352	11.88380751	19.94398114
01-07-2022 02:00	2.1755232	1.171428632	233.0435842	4.965645867	23.33662516	72.36183352	11.88380751	19.94398114
01-07-2022 03:00	2.1755232	1.171428632	233.0435842	4.965645867	23.33662516	72.36183352	11.88380751	19.94398114
01-07-2022 04:00	2.1755232	1.171428632	233.0435842	4.965645867	23.33662516	72.36183352	11.88380751	19.94398114
01-07-2022 05:00	2.1755232	1.171428632	233.0435842	4.965645867	23.33662516	72.36183352	11.88380751	19.94398114
01-07-2022 06:00	2.1755232	1.171428632	233.0435842	4.965645867	23.33662516	72.36183352	11.88380751	19.94398114
01-07-2022 07:00	2.1755232	1.171428632	233.0435842	4.965645867	23.33662516	72.36183352	11.88380751	19.94398114
01-07-2022 08:00	2.1755232	1.171428632	233.0435842	4.965645867	23.33662516	72.36183352	11.88380751	19.94398114

Table 3.1 - Extract of the dataset

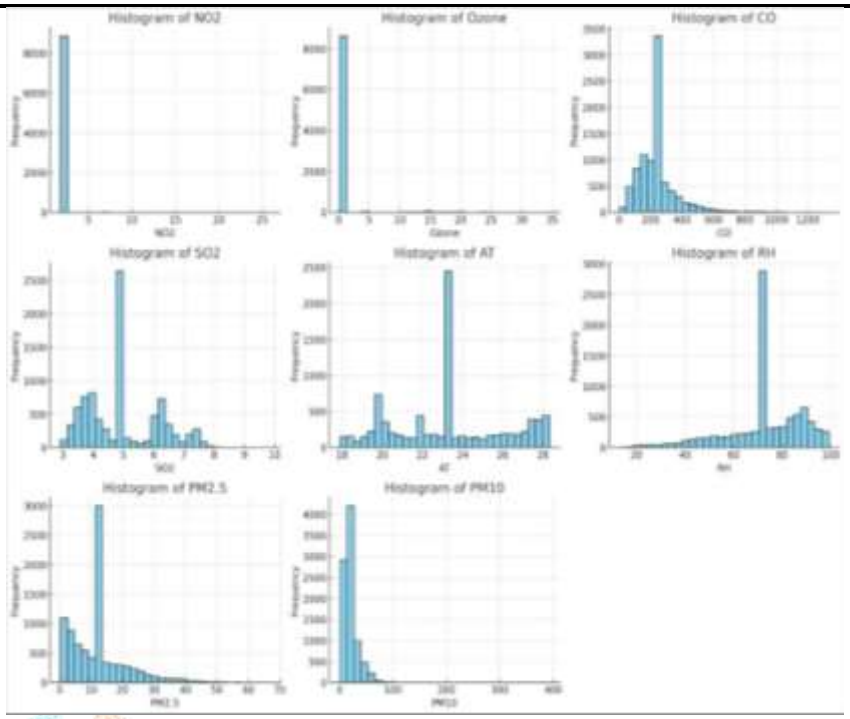

Figure 3.1 Histograms depicting the distribution of pollutants, ambient temperature (AT), and relative humidity (RH) over the study time period *Source: Authors' Inference*

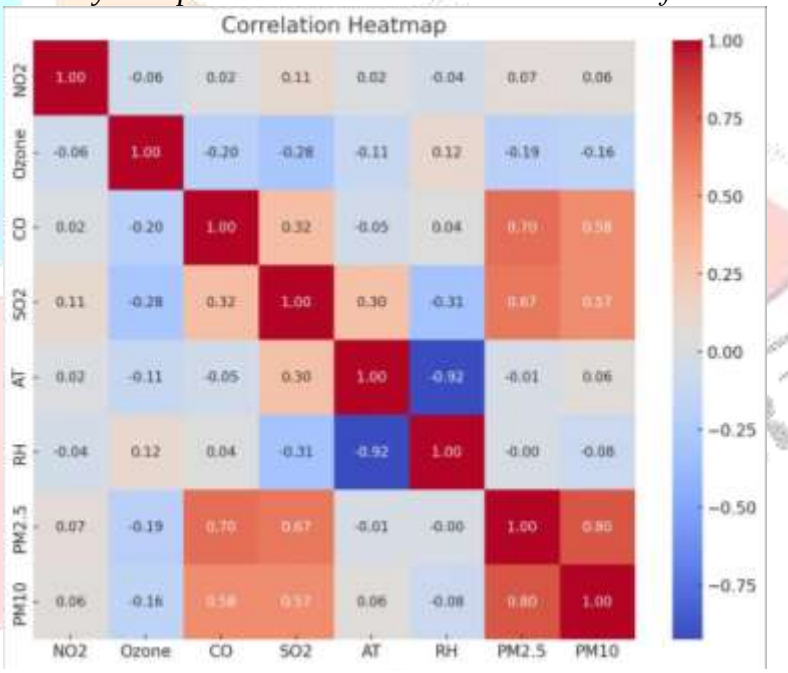

Figure 3.2 Correlation matrix depicting pairwise relationships among pollutant concentrations, AT, and RH. *Source: Authors' Inference*

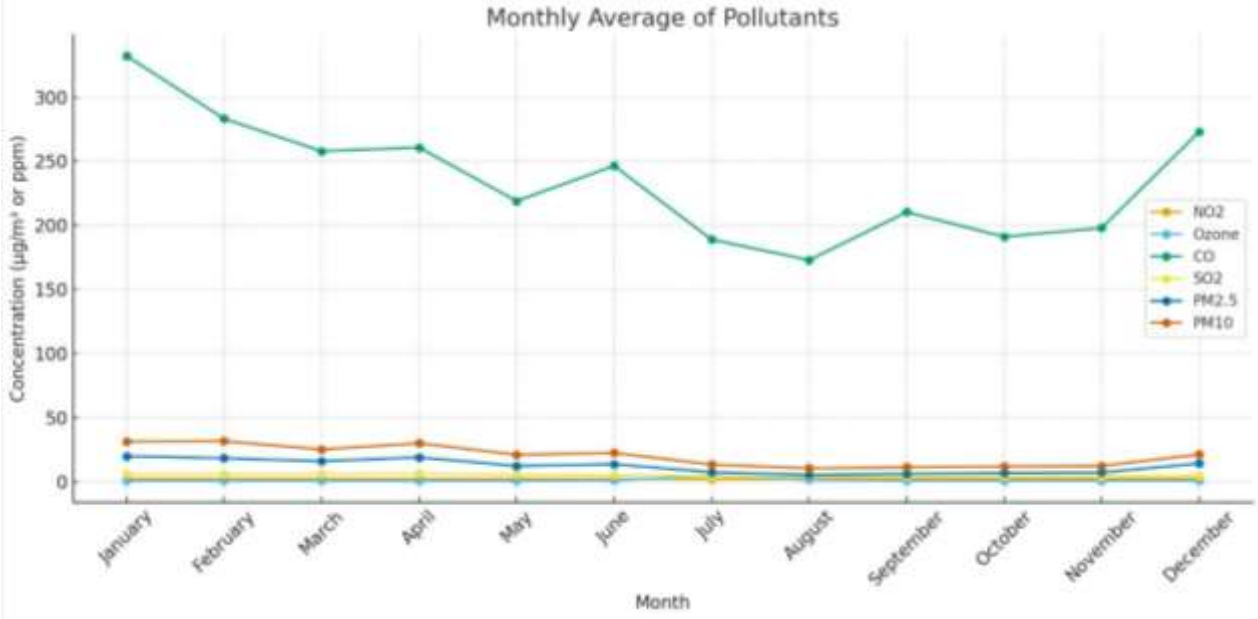

Figure 3.3 Time Series Plot showing monthly variation in average pollutant concentrations, illustrating seasonal trends across the twelve months. Source: Authors' Inference

3.2 Proposed System

A comprehensive IAQI model that computes sub-indices for particular pollutants in alignment with CPCB standards constitutes the basis of the system. In order to guarantee that the most important pollutant at any given time defines the risk level, the overall IAQI is subsequently obtained as the highest of these sub-indices. Forecasting is achieved through a lightweight model that projects IAQI values up to seven days ahead, based on recent observations and trend estimation. The predictions are classified into meaningful categories such as Good, Moderate, Poor, or Severe, accompanied by descriptive health advisories for interpretability.

3.3 Equations

The mathematical foundation of the model is summarized below.

The Indoor Air Quality Index (IAQI) for each pollutant is calculated using linear interpolation between CPCB breakpoints. For a pollutant concentration C_p, the IAQI is expressed as Eq.1

$$IAQI_p = \frac{(I_{high} - I_{low})}{(BP_{high} - BP_{low})} \times (C_p - BP_{low}) + I_{low} \quad ,$$

Eq.1

Where BP_{low} and BP_{high} are the lower and upper breakpoints related to the concentration C_p, while I_{low} and I_{high} are the corresponding IAQI values.

The overall IAQI at any time step is determined by taking the maximum of the sub-indices Eq.2 :

$$IAQI = max\{IAQI_{PM2.5},\ IAQI_{PM10},\ IAQI_{CO}\}.$$

Eq.2

For prediction, a short-term linear trend is applied to the recent IAQI values, with controlled noise introduced to reflect uncertainty Eq.3:

$$P_{t+1} = IAQI_{base} + \big(trend \times (t+1)\big) + \epsilon,$$

Eq.3

Where $IAQI_{base}$ the most recent index value is, the trend is the slope obtained from regression over the past 14 days, and it is a small random variation drawn from a normal distribution.

5. SYSTEM DESIGN

The design of the system follows a clear & concise plan. First, it takes a CSV file uploaded by users. It checks data to make sure key parts are there. It also fixes mix-ups in column names. Next, the work part uses CPCB rules to find IAQI for each bad air part. The top score sets the full IAQI for that time. Next, the forecasting stage uses statistical calculations to predict the IAQI values for the next seven days by extending the calculation. To ensure that the result is robust, residual correction is applied.
At last, as the end result, the values are displayed on a web page. The scores shown will have both raw IAQI numbers and type names. The tool is adaptive, so its access is available on both desktop and mobile. (*Figure 5.1)*

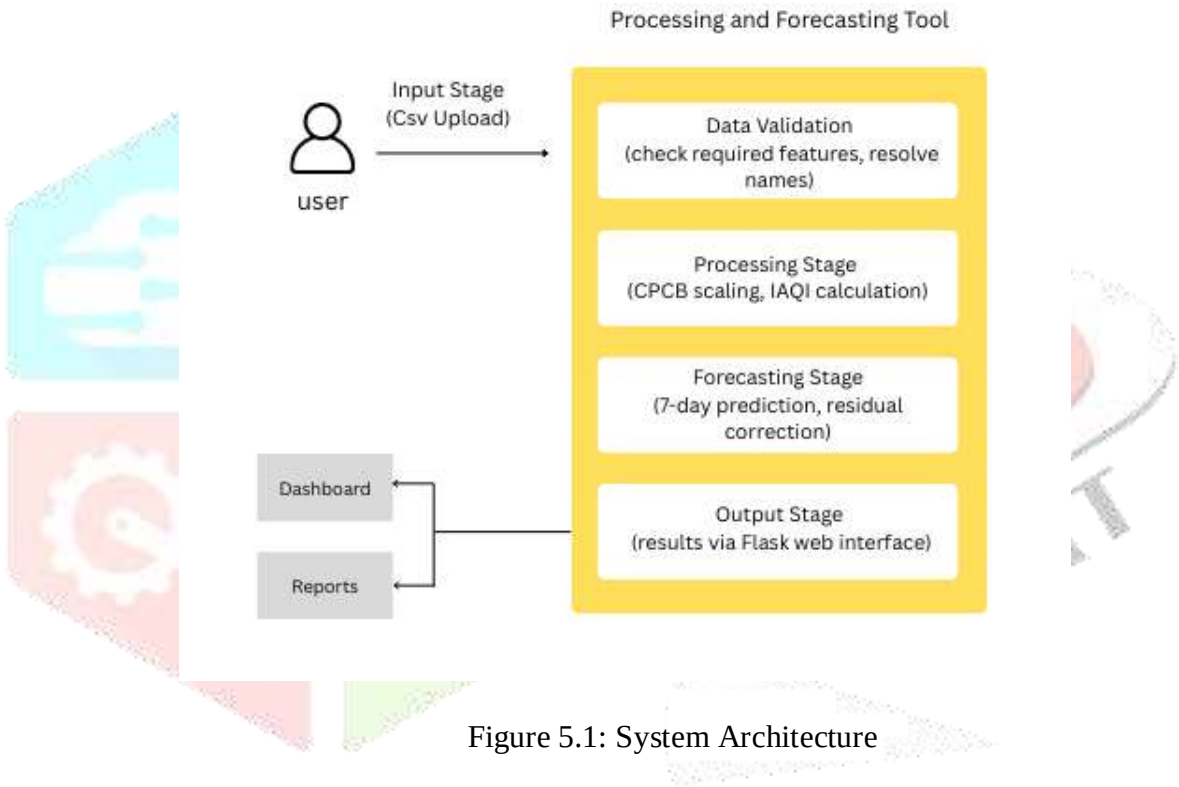

Figure 5.1: System Architecture

6. WORKING PRINCIPLE

The suggested system functions as an ongoing pipeline that converts unprocessed pollutant data into accurate forecasts of air quality. From the user's point of view, the procedure starts with something as easy as using the web interface to upload a CSV file. But before the uploaded dataset is prepared to produce predictions, it must go through a number of important processes behind the scenes. Ranging from upload validation to output processing the tool follows a number of crucial steps to make sure the the predicted data is accurate(*Figure 6.1)*.

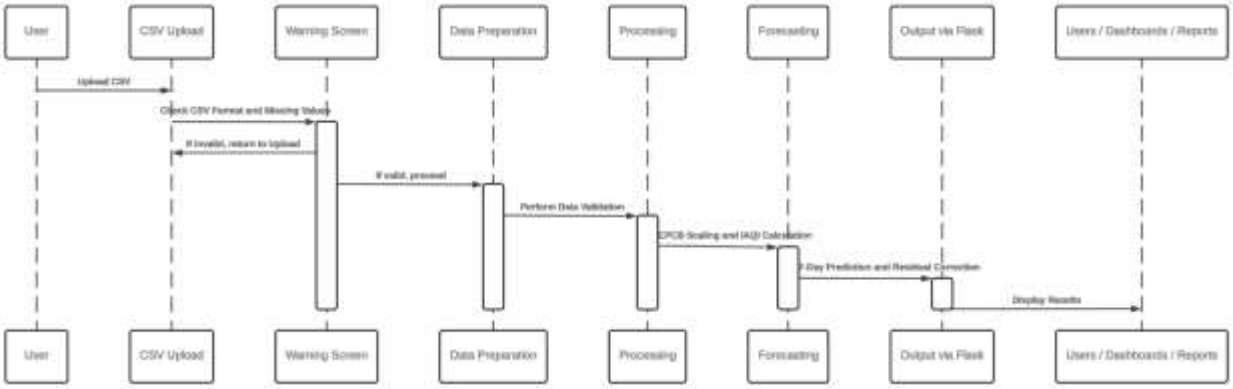

Figure 6.1: Workflow Diagram

Injection of data and validation are the initial steps(*Figure 6.2*). The key features, such as PM2.5, PM10, CO, temperature, and humidity, are available due to the fact that actual datasets may contain missing values, different column names, or abnormal formats. A back-end fuzzy matching is used to fix a misspelled column (such as “PM 2.5” instead of “pm25”). As a result, this allows the system to handle dirty databases that other, more rigid models may not be able to accommodate. Because of this, the system can tolerate flawed datasets that more strict models could otherwise fail.
The data gets cleaned and preprocessed post-validation. Extreme outliers are modified to come below reasonable environmental thresholds, and any values that are missing are assigned using statistical methods. This guarantees that inaccurate spikes or missing entries won't influence the pipeline's later stages. The next step is to compute the Indoor Air Quality Index (IAQI) after the dataset is in a suitable state. The raw levels of pollutants are plotted on an index scale of 0 to 500 in accordance with the established CPCB (Central Pollution Control Board) guidelines. The IAQI is computed independently for each pollutant, and the maximum of these is used to determine the aggregate IAQI for a specific time step. This effectively identifies the most important pollutant at that time.

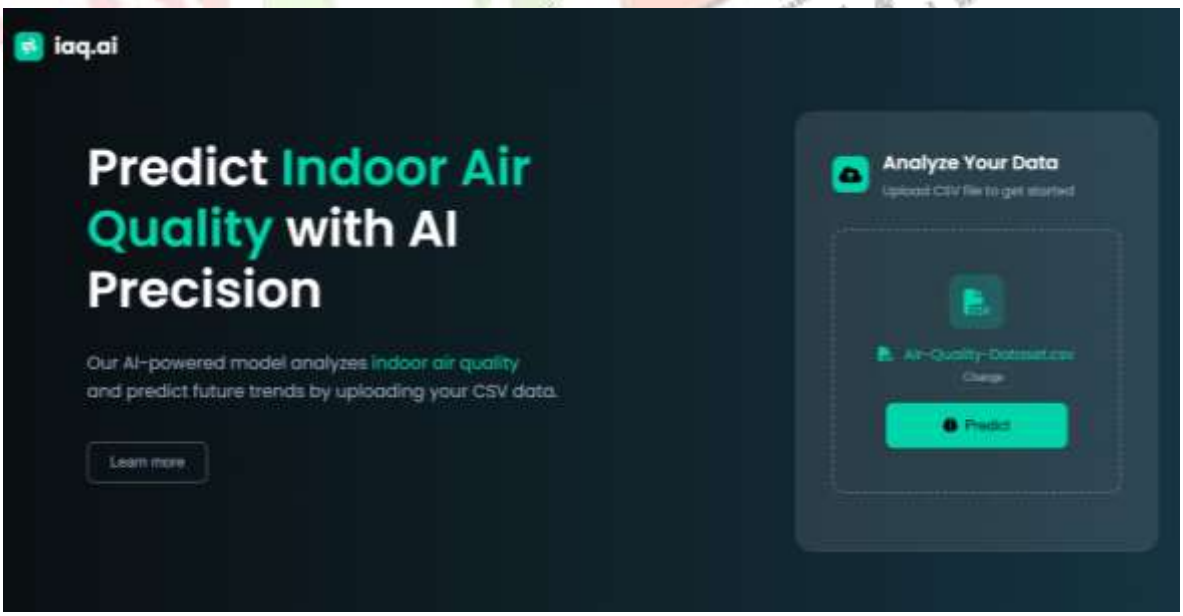

Figure 6.2: CSV Upload on Home Page

The prediction tool’s key objective is to make a future forecast. Avoiding just a present assessment of the air quality, the system forecasts IAQI values for the future using statistical approaches, typically up to seven days in advance. By analyzing past discrepancies between expected and observed values,

residual correction techniques improve future predictions and boost reliability. The predictions utilize just a small amount of controlled randomization to prevent excessively rigid linear forecasts that would overlook the natural variances seen in real air quality changes.

They are categorized into four qualitative categories—Good, Moderate, Poor, and Severe, after the IAQI results for the upcoming few days are obtained. These categories are connected with health warnings in order to ensure the findings are not just figures, but actionable information for households, businesses, and policymakers (*Figure 6.3*). The Flask-based web interface is then used to inform the user of the results. Interactive graphs that display the anticipated trends in addition to numerical numbers make the system both educational and easy to use on both desktop and mobile platforms.
The system's basic operation can be summed up as an orderly progression from unprocessed contamination measurements to knowledgeable forecasts: raw data is cleaned and computed into IAQI values, which are then presented as insightful information rather than raw datasets, for the general audience of the tool, who might not be technical experts, while also maintaining computational robustness.

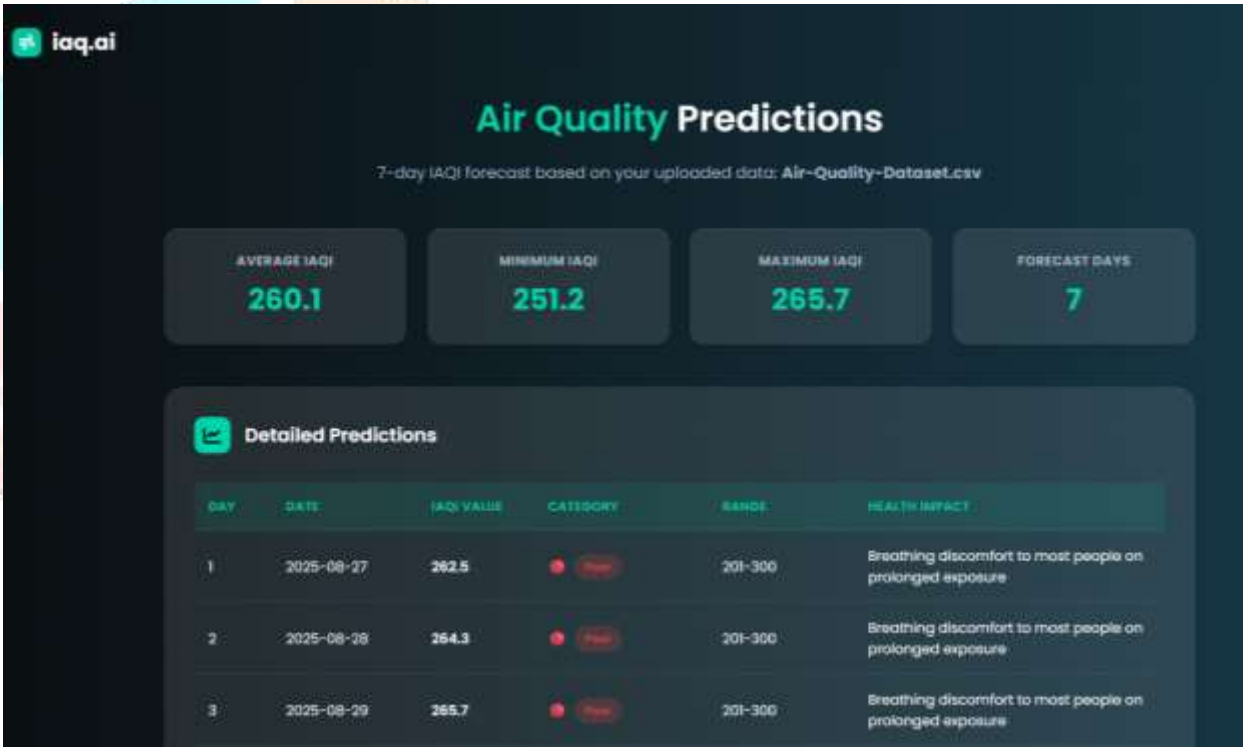

Figure 6.3: Result Screen With Predictions

7. CONCLUSION

This study ascertains that statistical prediction coupled with flexible web-service deployment is sufficient for reliable IAirQ prediction. By combining short-term forecasting with IAQI computation, this tool is an effort to narrow the distance existing between lab models and everyday use tools.
This is intelligence that thinks ahead, as opposed to our current IAQ systems, which simply report on pollution. The architecture is suitable for real-world applications where the quality of the data cannot be assured, and hence is robust to different datasets and can accommodate missing values and errors.
The significance of this system is not limited to academic illustration. It can also be useful for families to make judgments about the use of air purifiers and ventilation. It can assist occupational safety measures in the workplace. During times of high pollution risk, it can serve as an aid to decision-

making for policymakers, allowing for timely measures. Policymakers' decision-making ensures that insights are accessible to the general public rather than being restricted to technical settings.
However, there are still chances to improve even more. Future research might concentrate on:

- Longer-term forecasting (14–30 days, for example) with a combination of deep learning and statistical models.
- Instead of batch uploads, real-time interaction with IoT sensors allows for continuous live prediction.
- Wider coverage of pollutants, such as NO_2, ozone, and VOCs, for a more complete IAQ image.
- Personalized exposure modeling, in which forecasts take into account both personal health profiles and environmental data.
- Integration of a machine learning model through a pipeline into the website

One can hope that IAQ prediction will eventually develop as a universally trainable, trustworthy, and transparent technology. This piece is an effort to do that by straddling the line between stats and openness and accessibility. The frame of reference set out in this article allows individuals and institutions to politically deal with "the hidden enemies", or indoor pollution, as it turns into not merely a technical instrument inside the workplace but a vehicle to regulate overall standards of living.

8. ACKNOWLEDGMENT

This paper is supported by the CHRIST (Deemed to be University) Seed Money grant titled “Smart Application using IoT and AI towards Analysing Healthy and Sustainable Living Environment” (CU:CRP: SMSS-2349, Dated:11/01/2024)

9. REFERENCES

[1] World Health Organization. 2022. WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization, Geneva.
[2] Klepeis, N.E., Nelson, W.C., Ott, W.R., Robinson, J.P., Tsang, A.M., Switzer, P., Behar, J.V., Hern, S.C. & Engelmann, W.H. 2001. The National Human Activity Pattern Survey (NHAPS): A resource for assessing exposure to environmental pollutants. Journal of Exposure Analysis and Environmental Epidemiology, 11(3): 231–252.
[3] Fan, S., Hao, D., Feng, Y., Xia, K. & Yang, W. 2021. A hybrid model for air quality prediction based on data decomposition. Information, 12(5): 210.
[4] Zhao, L., Yang, Y. & Wu, Z. 2022. Review of communication technology in indoor air quality monitoring system and challenges. Electronics, 11(18): 2926.
[5] Gupta, A., Yadav, R. & Singh, P. 2021. Comparative study of machine learning models for air quality index prediction. Environmental Science and Pollution Research, 28: 56732–56745.
[6] Skiles, M.J., Lai, A.M., Olson, M.R., Schauer, J.J. & De Foy, B. 2018. Source apportionment of PM2.5 organic carbon in the San Joaquin Valley using monthly and daily observations and meteorological clustering. Environmental Pollution, 237: 366–376.
[7] Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C. & Baklanov, A. 2012. Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects. Atmospheric Environment, 60: 656–676.
[8] Li, Y. & Li, R. 2023. A hybrid model for daily AQI prediction and its performance during COVID-19 lockdown. Process Safety and Environmental Protection, 176: 673–684.
[9] Lu, W., Ai, T., Zhang, X. & He, Y. 2017. An interactive web mapping visualization of urban air quality monitoring data of China. Atmosphere, 8(8): 148.

[10] Brook, R.D., Franklin, B., Cascio, W., Hong, Y., Howard, G., Lipsett, M., ... & Smith, S.C. Jr. 2010. Air pollution and cardiovascular disease: A statement for healthcare professionals from the American Heart Association. Circulation, 121(21): 2331-2378.

[11] Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., ... & Dominici, F. 2017. Air pollution and mortality in the Medicare population. New England Journal of Medicine, 376(26): 2513-2522.

[12] Dominici, F., Peng, R.D., Barr, C.D., & Bell, M.L. 2010. Protecting human health from climate change: Preparing small islands for climate change. Bulletin of the World Health Organization, 88(12): 885-886.

[13] Jerrett, M., Burnett, R.T., Pope, C.A., Ito, K., Thurston, G., Krewski, D., ... & Thun, M.J. 2009. Long-term ozone exposure and mortality. New England Journal of Medicine, 360(11): 1085-1095.

[14] Kampa, M., & Castanas, E. 2008. Human health effects of air pollution. Environmental Pollution, 151(2): 362-367.

[15] Laden, F., Schwartz, J., & Speizer, F.E. 2006. Reduction in fine particulate air pollution and mortality: Extended analysis of the Harvard Six Cities Study. American Journal of Respiratory and Critical Care Medicine, 173(6): 667-672.

[16] Pope, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., & Thurston, G.D. 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Journal of the American Medical Association, 287(9): 1132-1141.

[17] Samoli, E., Peng, R.D., Ramsay, T., Pipikou, M., Touloumi, G., Dominici, F., ... & Katsouyanni, K. 2008. Acute effects of ambient particulate matter on mortality in Europe and North America: Results from the APHENA study. Environmental Health Perspectives, 116(11): 1480-1486.

[18] Woodruff, T.J., Grillo, J., & Schoendorf, K.C. 1997. The relationship between selected causes of postneonatal infant mortality and particulate air pollution in the United States. Environmental Health Perspectives, 105(6): 608-612.

[19] Zhang, Y., Liu, X., & Chen, W. 2019. Applications of machine learning methods in air quality monitoring and prediction: A review. Atmospheric Research, 224: 32-44.

[20] Zheng, Y., Zhang, Q., Tong, D., Davis, S.J., & Streets, D.G. 2018. Technology-driven interannual variability in global smog deaths. Nature Communications, 9(1): 1-9