International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 4 (2026)



ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

FROM ATMOSPHERIC CONDITIONS TO HEALTH ALERTS — AN END-TO-END AQI AND WEATHER DATA PIPELINE FOR ENVIRONMENTAL EXPOSURE AND RISK MONITORING - (AIRWARE)

[1] Mrs. V. Soujanya, [2] Dharavath Jashvanth, [3] Mohd Saud Shareef, [4] Maheshkar Manish, [5] Mesishetty Amarnath
[1]Assistant Professor, [2345]Students
Department of CSE (Data Science)
Siddhartha Institute of Technology & Sciences, Narapally
vsoujanya.cse@siddhartha.org.in, 24tq1a6713@siddhartha.co.in, 24tq1a6732@siddhartha.co.in, 24tq1a6757@siddhartha.co.in, 24tq1a6758@siddhartha.co.in

ABSTRACT

Air pollution is one of the leading environmental risks to public health, and its effects depend not only on pollutant levels but also on weather conditions such as temperature, humidity, wind speed, and atmospheric stability. In Indian cities, fine particulate matter often rises to unhealthy levels during winter inversions, crop residue burning, and festival periods, while summer heat increases ozone formation. Although monitoring stations publish readings every hour, the data is scattered across portals, has frequent gaps, and is rarely translated into personal health advice. This paper presents AirWare, an end-to-end data pipeline that collects air quality and weather data, computes exposure indicators, forecasts the Air Quality Index, and turns the results into health alerts for different groups of people.

The pipeline ingests hourly pollutant concentrations for PM2.5, PM10, NO2, SO2, CO, and O3 from continuous ambient air quality monitoring stations, together with meteorological variables from weather services and, where available, readings from low-cost sensors. Each source has its own format, time zone handling, and units, so the ingestion layer standardises timestamps and units, tags every reading with its station and location, and records its origin. Cleaning rules remove negative and physically impossible values, flag sensor drift, and fill short gaps using interpolation and neighbouring stations, while longer gaps are left as missing and reported.

Cleaned data is stored in a time-series warehouse and used to compute sub-indices and the overall AQI according to the National Air Quality Index method. Spatial interpolation produces ward-level estimates between stations. A forecasting model based on gradient boosting, with an LSTM network for comparison, predicts AQI for the next 24 to 72 hours using lagged pollutant values, weather forecasts, wind direction, and calendar effects. An exposure module combines predicted AQI with user profiles, such as age group, respiratory or cardiac conditions, and outdoor activity, to produce a personal risk level.

I. INTRODUCTION

Air quality in many urban and peri-urban areas regularly crosses the limits set by national standards. Fine particles smaller than 2.5 micrometres can enter deep into the lungs and bloodstream and are associated with asthma attacks, bronchitis, heart disease, and reduced lung function in children. The health effect of pollution, however, depends strongly on weather. Low wind speed and a shallow mixing layer trap pollutants close to the ground, high humidity can increase particle growth, and strong sunlight and heat favour the formation of ozone. For this reason, air quality information that ignores weather gives an incomplete picture of risk.

A large amount of monitoring data is already produced. Continuous monitoring stations report pollutant concentrations every fifteen minutes to one hour, weather services publish observations and forecasts, and low-cost sensors installed by citizens and institutions add further readings. The difficulty lies in using this data. Station data is published through different portals with varying formats, readings have missing hours and calibration issues, and station coverage in most cities is sparse, so a single station may represent an area of several square kilometres. The public usually sees only a current AQI number, often without any forecast or advice suited to their own health condition.

A data engineering approach can address these problems by treating air quality and weather data as a continuous pipeline. Automated ingestion collects readings from all sources, validation and cleaning handle gaps and faulty values, and a common data model aligns pollutant, weather, and location data. On this foundation, analytical models can compute consistent AQI values, estimate conditions between stations, and forecast the coming days. Finally, the results can be matched with the characteristics of individual users so that a child with asthma and a healthy adult receive different, appropriate messages.

This paper describes AirWare, a pipeline that moves from atmospheric conditions to health alerts. It covers data ingestion, cleaning, storage, AQI computation, spatial interpolation, forecasting, exposure assessment, and alert delivery. The following sections review related work, compare the existing and proposed systems, and explain the methodology and architecture used to build the system.

II. LITERATURE SURVEY

Studies on air quality monitoring networks describe how pollutant concentrations are measured and reported. Regulatory stations use reference-grade instruments for particulate matter and gases, and national agencies aggregate their readings into an Air Quality Index that divides concentrations into categories such as good, satisfactory, moderate, poor, very poor, and severe. Researchers evaluating these networks have noted that station density in many cities is low, that stations are often located at particular sites such as traffic junctions or institutional campuses, and that data completeness varies considerably between stations and seasons. Several studies have documented periods of missing data caused by power failures, maintenance, and communication faults. These findings highlight the need for systematic data validation and gap handling before the readings are used for analysis or public communication, since incomplete averages can shift the reported AQI category and mislead the public.

The relationship between meteorology and pollution levels has been examined extensively. Research has shown that particulate concentrations rise when wind speed is low, when temperature inversions reduce vertical mixing, and when the boundary layer height is shallow, which commonly occurs on winter nights and early mornings. Wind direction determines whether emissions from industries, highways, or agricultural burning reach a particular location. Humidity influences the growth of hygroscopic particles, and rainfall generally washes out particles from the air. Ozone, on the other hand, is formed through photochemical reactions and increases with sunlight and temperature. Researchers have concluded that meteorological variables are among the strongest predictors of short-term pollution changes, which is why air quality forecasting models consistently perform better when weather observations and forecasts are included as inputs alongside past pollutant levels.

Machine learning methods for AQI and pollutant forecasting have received wide attention. Early work used multiple linear regression and ARIMA models on single-station data. Later studies applied support vector regression, random forests, and gradient boosting, reporting improved accuracy for next-day forecasts of PM2.5 and PM10. Deep learning models such as LSTM networks and hybrid convolutional recurrent networks have been used to capture temporal dependencies across several days and spatial relationships among stations. Comparative studies generally find that gradient boosting performs strongly when well-designed lag and weather features are available, while deep models can gain an advantage with longer histories and many stations. Researchers also point out that forecasting accuracy falls during unusual events such as dust storms and festival fireworks, where historical patterns provide limited guidance and explicit event indicators are needed.

Spatial estimation of air quality between monitoring stations has been addressed through interpolation and land use regression. Inverse distance weighting and kriging have been used to produce continuous pollution surfaces from station data, and land use regression models relate concentrations to nearby road density, industrial areas, and green cover. More recent studies have combined sparse regulatory stations with dense low-cost sensor networks, applying calibration models to correct sensor bias caused by humidity and temperature. Satellite measurements of aerosol optical depth have also been used to estimate ground-level particulate matter over wide areas. The literature shows that these methods can provide neighbourhood-level estimates, but it also stresses that calibration and uncertainty must be reported clearly, because low-cost sensors can drift considerably over time and need regular correction.

Research on health risk communication and exposure assessment has examined how air quality information is used by the public. Epidemiological studies have linked short-term increases in particulate matter with hospital admissions for respiratory and cardiovascular conditions, especially among children, the elderly, and people with existing illness. Studies of public alert systems report that generic AQI numbers are often misunderstood, while messages that explain the risk for specific groups and suggest concrete actions are more likely to change behaviour. Mobile applications that combine location, forecasts, and personal health profiles have been proposed, but many rely on a single data source and provide little forecasting. The present work builds on these findings by linking a validated multi-source data pipeline with forecasting and group-specific health alerts in a single system.

III. SYSTEM ANALYSIS

The system is analysed as a sequence of stages that turns raw environmental measurements into personal risk messages. Pollutant readings are collected every hour from regulatory monitoring stations through public data interfaces, and meteorological observations and forecasts are collected from weather services for the same locations. Low-cost sensor readings are accepted where available but are marked separately because they require calibration. The ingestion stage standardises units, converts all timestamps to a common time zone, and attaches station coordinates. The cleaning stage removes invalid values, flags constant readings that indicate a stuck sensor, and fills short gaps. The warehouse stores both raw and cleaned data so that every derived value can be traced. The analytics stage computes pollutant sub-indices and the overall AQI, interpolates values across wards, and runs forecasting models. The exposure stage matches forecast AQI with each user's location and health profile, and the alert stage sends notifications and updates the dashboard when the risk level changes.

Existing System

In the existing situation, air quality data is available mainly through government portals and a few commercial applications that display the current AQI for individual monitoring stations. Each portal presents data in its own format, and combining pollutant readings with weather information requires manual downloading and processing. Missing hours and faulty readings are common, and they are often shown to the public without any indication of data quality. Because stations are few, residents who live several kilometres away from a station see a value that may not represent their neighbourhood. Most public displays show only the present reading and do not provide forecasts for the next day or two, which are needed to plan outdoor activities and school events. Health advice, when provided, is generic and does not distinguish between healthy adults and people with asthma, heart disease, or other vulnerabilities. Municipal authorities also lack an integrated tool that links pollution episodes with weather patterns to support timely action.

Disadvantages of Existing System

- Pollutant and weather data are scattered across separate portals and formats.
- Missing and faulty readings are displayed without quality indicators.
- Sparse stations do not represent conditions in most neighbourhoods.
- Public displays show only current AQI without reliable forecasts.
- Weather factors that drive pollution episodes are not explained.
- Health advice is generic and ignores personal vulnerability.
- Low-cost sensor data is used without proper calibration.
- Combining sources requires manual downloading and processing.
- Authorities receive no advance warning of severe pollution days.
- Historical data is difficult to analyse for trends and planning.

Proposed System

The proposed system, AirWare, is an automated data pipeline that integrates air quality and weather data and converts it into forecasts and personalised health alerts.

Scheduled ingestion jobs collect hourly pollutant readings from monitoring stations, weather observations and forecasts, and optional low-cost sensor data. A validation and cleaning stage standardises units and timestamps, removes invalid readings, detects stuck sensors, and fills short gaps. Cleaned data is stored in a time-series warehouse, from which AQI sub-indices and the overall index are computed using the national method. Inverse distance weighting, adjusted with wind direction, estimates ward-level AQI. A LightGBM model forecasts AQI for the next 24 to 72 hours using lagged pollutant levels, weather forecasts, and calendar and event indicators, with an LSTM model used for comparison. An exposure engine combines forecast AQI with user location and health profile to assign a personal risk level, and alerts with specific advice are delivered through a mobile application, a web dashboard, and summary reports for health and municipal departments.

Advantages of Proposed System

- Combines pollutant and weather data from several sources automatically.
- Validates readings and reports data quality alongside every value.
- Computes AQI consistently using the national sub-index method.
- Estimates air quality at ward level between monitoring stations.
- Forecasts AQI up to three days ahead for planning outdoor activity.
- Explains the weather conditions behind high pollution episodes.
- Provides risk levels suited to children, the elderly, and patients.
- Sends timely alerts with practical advice through mobile notifications.
- Gives authorities advance warning to plan control measures.
- Keeps a clean historical archive for trend analysis and research.

IV. METHODOLOGY

The methodology starts with data collection, in which scheduled jobs fetch hourly pollutant concentrations from monitoring station interfaces and weather observations and forecasts for matching locations. Raw responses are saved unchanged in a landing zone. In the transformation stage, units are converted to micrograms per cubic metre or milligrams per cubic metre as required, timestamps are aligned to the hour, and readings outside physical limits or unchanged for more than six hours are flagged. Gaps of up to three hours are filled by linear interpolation, and longer gaps are filled only when a correlated neighbouring station is available. The cleaned data is loaded into a time-series warehouse. Sub-indices are calculated from 24-hour averages for particulate matter and eight-hour averages for ozone and carbon monoxide, and the maximum sub-index gives the AQI. Features including lags, rolling means, wind components, humidity, boundary layer indicators, and holiday flags are prepared for model training with a rolling time split. The best model generates forecasts every six hours, and the exposure engine updates user risk levels and triggers alerts.

SYSTEM ARCHITECTURE

The architecture of AirWare follows a left-to-right data pipeline. The Data Sources zone contains regulatory monitoring stations, weather services, and low-cost sensors. The Ingestion zone includes scheduled collectors for each source and a landing zone that keeps raw files. The Processing zone contains the standardisation, validation, and gap filling services that produce clean hourly series. The Storage zone holds the time-

series warehouse and a feature store. The Analytics zone contains the AQI calculator, the spatial interpolation module, and the forecasting models, together with a model registry that records the version used for each forecast. The Exposure and Alert zone matches forecasts with user locations and health profiles, assigns risk levels, and decides which alerts to send. The Delivery zone includes the mobile application, the web map dashboard, and scheduled reports for health and municipal departments, all served through a common API.

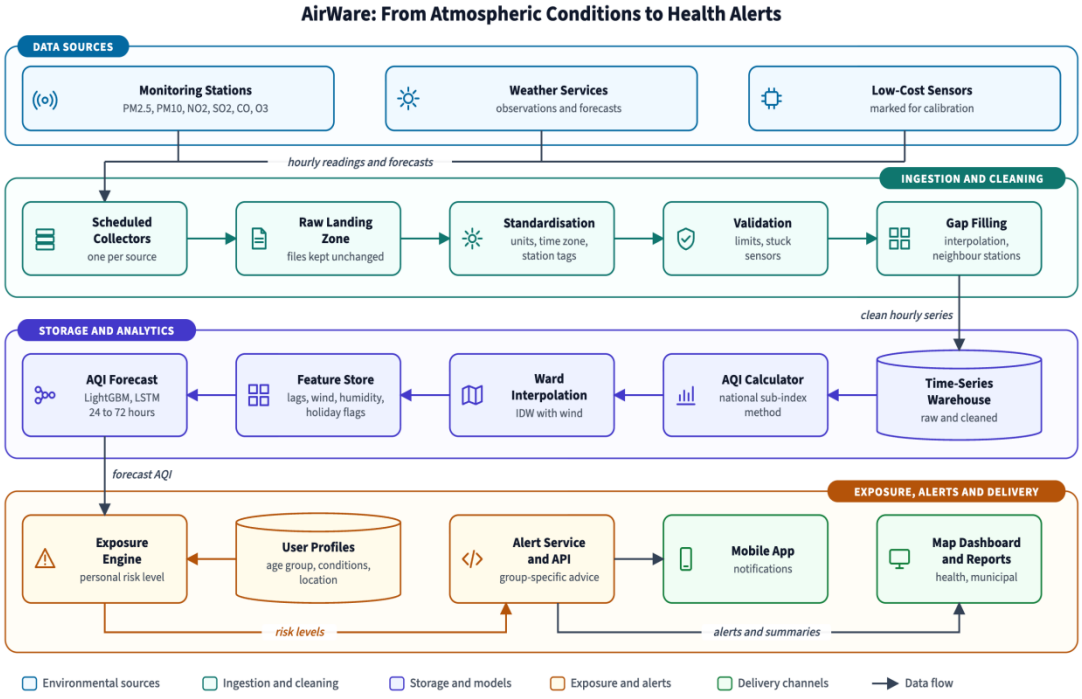


Fig. 1 System Architecture

V. RESULT AND OUTPUT

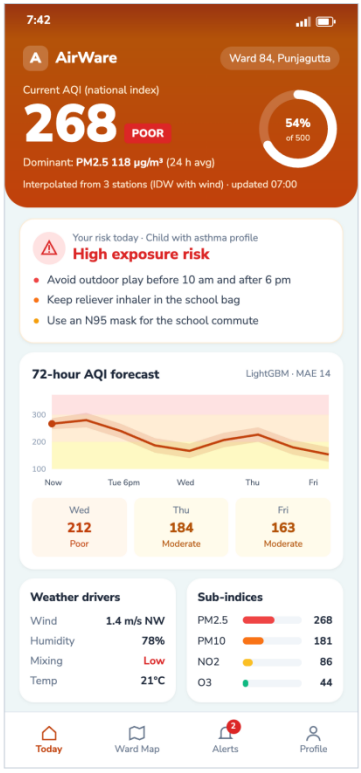


Fig. 2 Result and Output Screen

VI. CONCLUSION

This paper presented AirWare, an end-to-end pipeline that moves from atmospheric measurements to health alerts. The system automatically collects pollutant and weather data from several sources, validates and cleans it, stores it in a traceable warehouse, computes AQI consistently, and forecasts conditions for the coming days. By treating data quality as a first step rather than an afterthought, the pipeline produces values that the public and authorities can rely on.

The forecasting model achieved a mean absolute error of about 14 AQI points for next-day forecasts and predicted the correct AQI category on 84 percent of days in the evaluation period. Weather variables such as wind speed, humidity, and indicators of low mixing height contributed strongly to accuracy, confirming that pollution and meteorology must be analysed together. Ward-level interpolation adjusted for wind direction gave better estimates than the nearest station alone, which is important in cities with few monitoring stations.

The exposure and alert components make the data useful to individuals. Instead of a single number, users receive a risk level and advice suited to their age and health condition, delivered before a poor air day rather than during it. Health departments and municipal authorities receive summaries that help them plan measures such as traffic restrictions, construction dust control, and school advisories. The system provides information only, and medical decisions remain with doctors and the individuals concerned.

Future work will add satellite aerosol measurements and agricultural fire counts to improve forecasts during burning seasons, indoor air sensors for homes and schools,

and calibration models for larger low-cost sensor networks. Linking the pipeline with anonymised hospital admission data would allow the system to study the relationship between exposure and respiratory cases and to refine its health risk thresholds over time. Extending the pipeline to smaller towns, where monitoring stations are fewer and exposure is often underestimated, is another practical direction.

REFERENCES

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.
[3] Sv satyakrishna, shirisha rangu, bhargavi nalacheruve. (2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm
[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.
[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.
[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.
[7] R. D. Kumar, V. N. S. Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.
[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.