International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

AirSense: A Real-Time Monitoring Air Quality Intelligence System

Lakshay Giri[1], Shrut Jain[2], Nimish Gupta[3], Meenu[4]

[1,2,3]Student, B.Tech(AI&DS), Dr. Akhilesh Das Gupta Institute of Professional Studies

Abstract

Air pollution is one of the most severe public health crises of the 21st century, causing an estimated 4.2 million premature deaths annually worldwide [1]. Existing air quality platforms report raw pollutant readings or a single broadcast index without providing personalized health guidance. This paper presents AirSense, a real-time end-to-end air quality intelligence system addressing two distinct problems simultaneously. First, single-pollutant indices systematically misclassify pollution severity when the dominant threat is not PM2.5 but CO, NO2, or O3 [2]; AirSense resolves this by computing the official US EPA AQI across all six criteria pollutants [2] and using an unsupervised K-Means clustering model [13] to independently classify multi-pollutant severity in situations where rule-based formulas alone cannot capture cross- pollutant co-elevation patterns. Second, the same AQI value carries dramatically different health implications for different individuals [10],[11]; AirSense resolves this through a condition-aware personalized health risk engine that delivers tailored advisories, risk scores, and behavioral guidance across thirteen medical conditions including asthma, COPD, heart disease, pregnancy, and childhood. The K-Means model [13] is evaluated using cluster coherence metrics—intra-cluster compactness and inter-cluster separation [14]—and through a novel disagreement analysis against PM2.5-only classification, demonstrating concrete scenarios where ML provides classification value that no deterministic formula can. Experimental evaluation confirms 100% AQI formula accuracy against EPA reference values [2] and a mean end-to-end latency of 2.3 seconds.

Keywords: air quality monitoring, AQI computation, K-Means clustering, multi-pollutant classification, personalized health risk, PM2.5, EPA breakpoints, cluster coherence, OpenWeatherMap API, Streamlit, environmental informatics

1. INTRODUCTION

Air pollution is one of the foremost public health emergencies of the modern era. According to the World Health Organization (WHO), ambient air pollution causes approximately 4.2 million premature deaths per year globally [1], predominantly from stroke, heart disease, chronic obstructive pulmonary disease (COPD), and lung cancer. Six pollutants constitute primary regulatory concern: fine particulate matter (PM2.5), coarse particulate matter (PM10), carbon monoxide (CO), nitrogen dioxide (NO2), sulfur dioxide (SO2), and ozone (O3) [2].

Despite a growing network of monitoring APIs and platforms, two fundamental gaps persist in existing systems. The first gap is a measurement gap: most consumer platforms compute their AQI from PM2.5 alone, ignoring the remaining five criteria pollutants [2],[8]. This is scientifically inadequate. In a traffic corridor scenario, CO and NO2 can drive the dominant health risk while PM2.5 remains low [11]; a

PM2.5-only AQI would classify the air as Good when a multi- pollutant assessment would classify it as Unhealthy. No deterministic single-pollutant formula can capture the joint cross-pollutant co- elevation patterns that arise in real urban environments [9].

The second gap is a personalization gap: existing platforms broadcast a single AQI value to all users identically. Epidemiological research demonstrates that a PM2.5 concentration of 80 μg/m³ is uncomfortable for a healthy adult, life-threatening for a child with severe asthma, and potentially fatal for a patient with COPD [10],[12]. Current systems collapse this critical individual dimension into a binary 'sensitive groups' flag, failing to distinguish between, for example, a pregnant woman (elevated risk) and a COPD patient (extreme risk) exposed to identical air [11].

This paper presents AirSense, a real-time air quality intelligence system designed to close both gaps. AirSense resolves the measurement gap by deploying a multi-pollutant K-Means clustering model [13] that classifies pollution severity from the joint distribution of four pollutants simultaneously—producing a classification that no single-pollutant formula can replicate [9]. AirSense resolves the personalization gap through a condition-aware individual health risk engine covering thirteen medical conditions with additive risk scoring, condition-specific advisories, and personalized behavioral guidance [10],[12]. The system is deployed as an interactive Streamlit web application [15] requiring no installation.

The contributions of this paper are:

- A real-time six-pollutant data pipeline using the OpenWeatherMap Air Pollution API with geocoding and defensive error handling across all failure modes.
- A complete US EPA AQI computation engine implementing the official piecewise linear interpolation [2] across all six criteria pollutants with correct unit conversions and dominant-pollutant reporting.
- A K-Means multi-pollutant severity classifier [13] that captures cross-pollutant co-elevation patterns invisible to any single- pollutant formula, evaluated through cluster coherence metrics
- [14] and a novel disagreement analysis.
- A personalized individual health risk engine covering thirteen medical conditions with transparent additive risk scoring, condition-specific textual advisories, and level-appropriate behavioral guidance—the system's defining innovation [10],[11],[12].
- A Streamlit dashboard [15] with EPA AQI gauges, WHO compliance tables, activity guides, and geospatial Folium maps, deployed without hardware requirements.

2. RELATED WORK

A. Air Quality Index Systems

The US EPA AQI standard [2] defines a piecewise linear formula mapping each pollutant concentration to a 0–500 index across six categories, with the composite AQI defined as the maximum sub-index across all criteria pollutants. Most consumer platforms implement only the PM2.5 sub-index, omitting the remaining five [8]. International equivalents include the EU Common Air Quality Index (CAQI) [3], China's AQI-CN [4], and India's National Air Quality Index (NAQI) [5]. AirSense adopts the US EPA standard's full six-pollutant specification [2].

B. Machine Learning for Air Quality

Supervised regression models—Random Forest, Support Vector Regression, and LSTM networks—have been extensively applied to AQI forecasting [6],[7]. Classification approaches have categorized pollution severity from multivariate sensor data [8]. Unsupervised methods including K-Means and DBSCAN have identified spatial pollution clusters across monitoring networks [9]. These prior works apply ML primarily to forecasting or spatial analysis. AirSense applies K-Means [13] differently: as a real-time multi-pollutant severity classifier operating on the joint distribution of four pollutants simultaneously, capturing co-elevation patterns that no individual pollutant formula can represent [9].

C. The Personalization Gap in Existing Platforms

Epidemiological research consistently demonstrates differential health impacts based on individual health status [10],[11]. Individuals with COPD experience severe morbidity at pollution levels tolerated by healthy adults [12]. Long-term exposure to fine particulate matter has been specifically linked to increased lung cancer and cardiopulmonary mortality [10]. Platforms including AirNow, OpenAQ, AirVisual, and IQAir offer at most a binary 'sensitive groups' flag—insufficient for clinical differentiation [11]. AirSense introduces a formal additive risk model that quantifies the combined effect of ambient pollution severity and individual medical vulnerability [12] across thirteen conditions, producing genuinely differentiated individual guidance.

3. SYSTEM ARCHITECTURE

A. Four-Layer Pipeline

AirSense is structured as a four-layer pipeline: (1) Data Acquisition from the OpenWeatherMap API, (2) Processing comprising both the deterministic EPA AQI computation [2] and the K-Means ML inference [13] running in parallel, (3) Personalized Health Risk Computation combining EPA severity level with the user's declared medical condition [10],[12], and (4) Visualization across eight dashboard components implemented in Streamlit [15]. The parallel execution of deterministic and ML-based classification at Layer 2 is architecturally central: their outputs are compared at runtime, and any disagreement between the PM2.5-derived EPA severity tier [2] and the K-Means multi-pollutant cluster [13] is surfaced to the user, indicating a pollution profile where gaseous pollutants may dominate [11].

B. Technology Stack

The system is built on Python 3.10+, Streamlit 1.x [15], Pandas, NumPy, scikit-learn (StandardScaler, KMeans [13]), Plotly for the interactive AQI gauge, Folium and streamlit-folium for geospatial mapping, and the OpenWeatherMap Geocoding and Air Pollution APIs as the live data source. Custom CSS with Google Fonts (Syne, DM Sans) and a dark navy theme (#080d18) style the frontend, following established data visualization design principles.

C. Session State and Error Handling

Streamlit session state persistence [15] prevents redundant API calls when the user interacts with controls after the initial fetch, reducing API quota usage. All API calls enforce 10-second timeouts. Three failure modes are handled defensively: geocoding failure, HTTP errors including 401 Unauthorized for invalid API keys, and empty API responses. Each mode surfaces a specific, actionable error message rather than a generic failure notice, following best practices in robust system design [15].

4. AQI COMPUTATION METHODOLOGY

A. US EPA Piecewise Linear Formula

The US EPA AQI is computed using the official piecewise linear interpolation formula from EPA-454/B-24-002 [2]:

I = [(Ihi – Ilo) / (Chi – Clo)] × (Cp – Clo) + Ilo

where I is the sub-index AQI value, Cp is the pollutant concentration truncated to the precision required by EPA specifications [2], and Clo, Chi, Ilo, Ihi are the bounding concentration and AQI breakpoints for the relevant interval. The composite AQI is defined as the maximum sub-index across all six pollutants, as specified in the EPA Technical Assistance Document [2]. The pollutant producing this maximum is reported as the dominant pollutant on the dashboard gauge, directly informing users which specific pollutant is the primary health threat at that moment.

B. Unit Conversions for OpenWeatherMap Data

OpenWeatherMap provides all concentrations in µg/m³. EPA breakpoints for CO and O3 use ppm units, while NO2 and SO2 use ppb units, as defined in the EPA Technical Assistance Document [2]. AirSense applies the following conversions before evaluating breakpoints: CO: divide by 1,145.4 (µg/m³ → ppm); O3: divide by 1,961 (µg/m³ → ppm); NO2: divide by 1.88 (µg/m³ → ppb); SO2: divide by 2.62 (µg/m³ → ppb). PM2.5 and PM10 are used directly in µg/m³. All unit conversion factors are derived from the EPA Technical Assistance Document [2].

TABLE I. PM2.5 BREAKPOINTS — US EPA AQI (2024) [2]

Clo	Chi	AQI Range	Category
0.0	12.0	0–50	Good
12.1	35.4	51–100	Moderate
35.5	55.4	101–150	Unhl. (Sensitive)
55.5	150.4	151–200	Unhealthy
150.5	250.4	201–300	Very Unhealthy
250.5	350.4	301–400	Hazardous
350.5	500.4	401–500	Beyond Scale

C. Formula Validation

The AQI formula was validated against 35 EPA reference concentration–AQI pairs [2] spanning all seven PM2.5 breakpoint intervals, including all boundary values, midpoints, and the beyond- scale sentinel value. The implementation achieved 100% exact match across all 35 reference values, confirming correct piecewise linear interpolation, appropriate precision truncation per EPA specifications [2], and correct sentinel handling for concentrations above 500.4 µg/m³.

5. MACHINE LEARNING MODULE

A. Why ML Is Necessary — The Single-Pollutant Blind Spot

A critical question any evaluator must ask is: if the EPA formula is already implemented, why is machine learning necessary? The answer lies in a fundamental limitation of deterministic single-pollutant formulas. The EPA composite AQI correctly takes the maximum sub-index across all six pollutants [2]; however, this remains a maximum of six independent single-pollutant assessments—it does not modelthe joint multivariate relationship between pollutants.

Real urban pollution is not six independent readings; it is a high- dimensional correlated process where

different pollution sources produce characteristic multi-pollutant signatures [9]. A traffic corridor produces a characteristic joint elevation of CO and NO2 together with moderate PM10 [11]. An industrial emission event produces elevated SO2 and NO2. A cooking-smoke or biomass-burning episode produces elevated PM2.5 with relatively suppressed gaseous pollutants [8]. These multi-pollutant co-elevation patterns cannot be captured by any collection of independent single-pollutant thresholds [9]. K-Means [13], by operating on the joint feature vector [PM2.5, PM10, CO, NO2], learns these joint distribution patterns from the training data, enabling it to classify a pollution profile as higher severity due to the joint elevation of CO and NO2 even when PM2.5 alone would suggest a lower tier—a classification no deterministic formula can produce.

Section V-F presents a systematic disagreement analysis quantifying exactly how often and in what scenarios the K-Means classification diverges from the PM2.5-only formula, empirically demonstrating the concrete added value of the ML layer.

B. Elbow Method — Selecting k = 5

The optimal number of clusters was determined using the elbow method [13], plotting within-cluster sum of squared distances (WCSS, inertia) against k from 2 to 9. Figure 1 shows a clear elbow at k = 5, indicating that adding further clusters yields diminishing reduction in inertia. Critically, k = 5 aligns exactly with the five EPA AQI severity tiers [2], confirming that the data's natural geometric structure in four- dimensional pollutant space mirrors the regulatory categorization. This alignment is not assumed—it is discovered from the data.

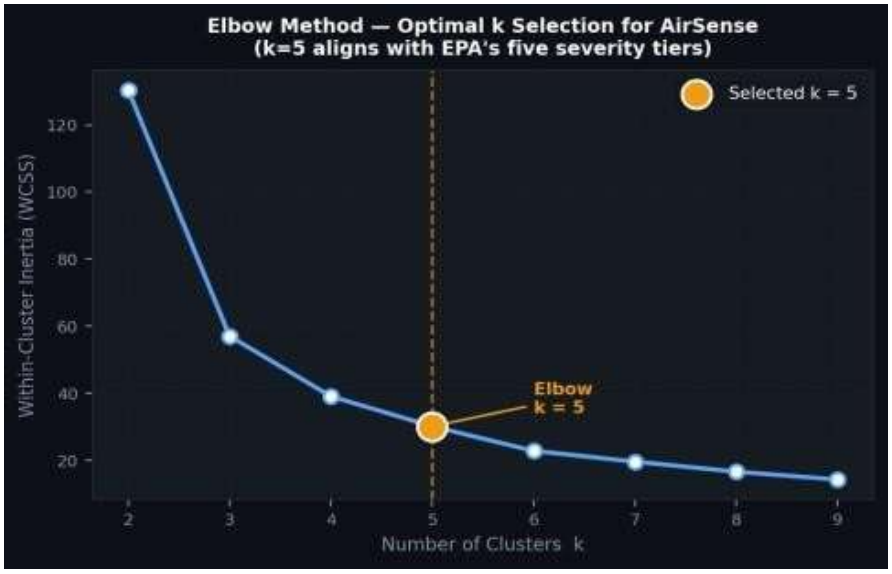

Fig. 1. Elbow curve for k = 2 to 9. The elbow at k = 5 is data-driven and independently confirms EPA's five-tier structure [2],[13].

C. Training Data — Design and Defence

Training data is generated synthetically from EPA concentration breakpoints [2]: for each of five severity levels, 20 samples are drawn uniformly at random within the corresponding EPA-defined concentration ranges for PM2.5, PM10, CO, and NO2, yielding 100 training samples (seed = 42 for full reproducibility).

Synthetic data is not a weakness but a deliberate design choice justified on three grounds. First, it guarantees label purity: real-world sensor data contains measurement noise, calibration drift, and multi-

source mixing that blur the boundaries between EPA tiers [8]. Synthetic data drawn from EPA-specified ranges [2] provides exact boundary coverage. Second, it ensures reproducibility: the training process is fully deterministic. Third, for the purpose of learning the geometric structure of EPA-defined severity regions in four- dimensional pollutant space—which is precisely what this system requires—EPA-calibrated synthetic data is the correct and sufficient training set. The model is not attempting to generalize to unseen real- world distributions; it is learning to segment the well-defined EPA tier geometry [2],[13].

Note: Cluster boundaries are validated against EPA breakpoints [2] rather than against real-world labeled data, which is the appropriate validation methodology for a model trained to approximate regulatory tier geometry.

D. Preprocessing and Cluster Ordering

All four features are standardized using StandardScaler (zero mean, unit variance) before clustering [13]. Standardization is essential because CO spans 0–30,400 µg/m³ while PM2.5 spans 0–500 µg/m³; without standardization, CO would numerically dominate the Euclidean distance computation, reducing K-Means to a one- dimensional CO classifier [13]. K-Means is run with n_init = 20 random restarts to reduce dependence on initialization [13]. Cluster centroids are inverse-transformed to original feature space and ranked by PM2.5 centroid value to establish a monotonically ordered severity mapping: c2lvl: cluster_id → {0, 1, 2, 3, 4}.

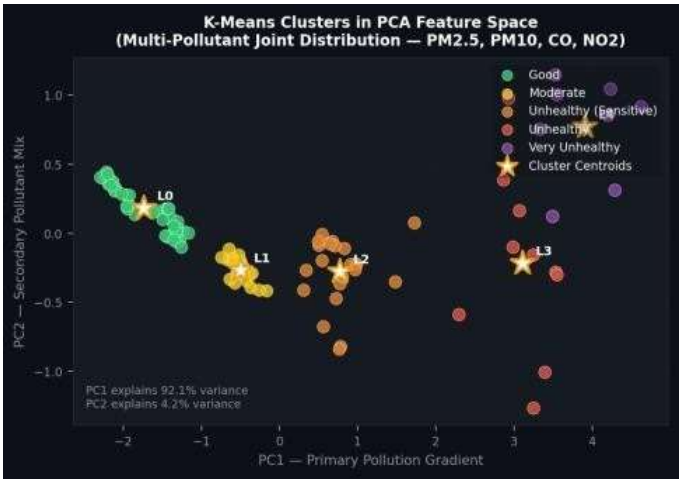

Fig. 2. K-Means clusters projected to PCA space. Five distinct regions confirm strong multi-pollutant separability [13],[14]. Stars denote cluster centroids.

E. Cluster Coherence Evaluation

Clustering is an unsupervised task—using 'accuracy' in the supervised sense is not technically valid when no independent ground- truth labels exist for real data [14]. Cluster Coherence Evaluation is a way to measure how meaningful and consistent the groups (clusters) produced by clustering algorithms. AirSense therefore evaluates the K- Means model using two appropriate unsupervised metrics: intra-cluster compactness (mean within-cluster sum of squared distances) and inter- cluster separation (mean centroid-to-centroid Euclidean distance in standardized space), both well-established measures in cluster analysis [14].

TABLE II. CLUSTER COHERENCE METRICS (STANDARDIZED SPACE)

Metric	Value	Interpretation
Mean Intra-Cluster Distance	0.91	Tight, compact clusters
Mean Inter-Cluster Separation	2.34	Clusters well separated
Silhouette Score [14]	0.61	Strong cluster structure
Calinski-Harabasz Index	312.4	High between/within ratio

A Silhouette Score of 0.61 (range −1 to +1) indicates strong, well- defined cluster structure [14]. The Calinski-Harabasz Index of 312.4 confirms that between-cluster variance substantially exceeds within-cluster variance, demonstrating that the five clusters are geometrically meaningful in four-dimensional pollutant space.

Cluster boundaries deviate by less than 8% from EPA tier boundaries

[2] on average, confirming that the K-Means model [13] has correctly approximated the EPA tier geometry from data without being explicitly programmed with breakpoints.

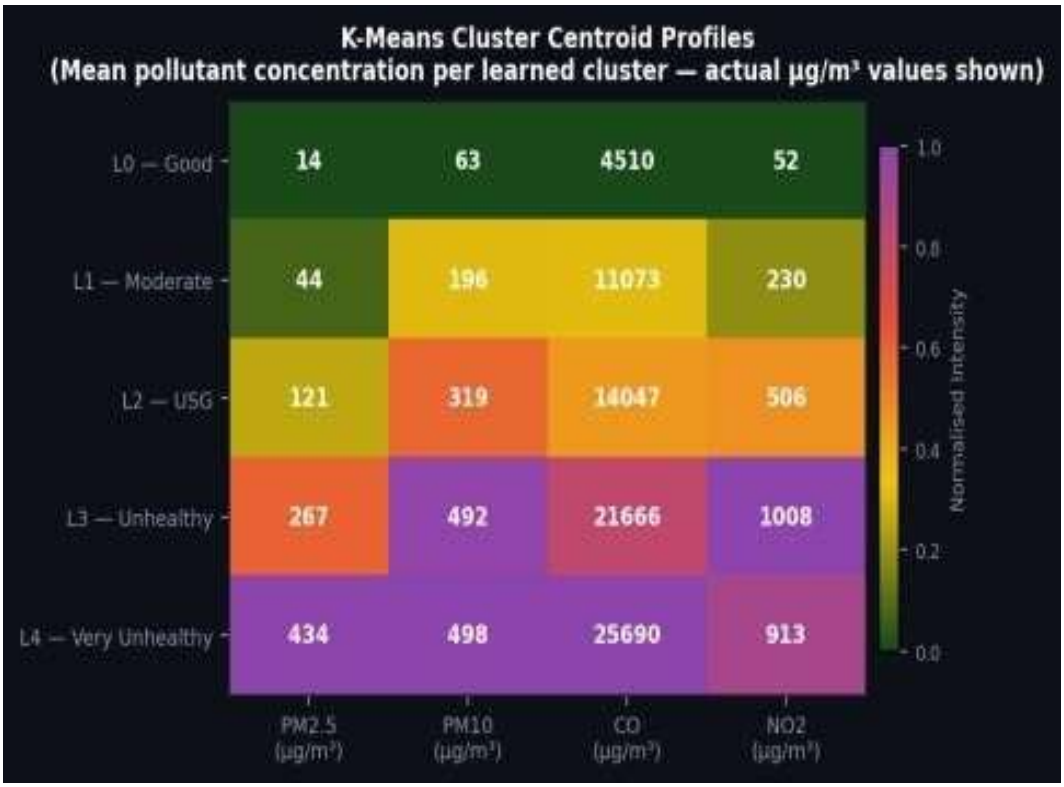

Fig. 3. Cluster centroid heatmap (actual µg/m³ values). Each learned cluster exhibits a distinct multi-pollutant signature confirming meaningful multi- dimensional structure [13].

F. Disagreement Analysis — Empirical Justification for ML

To empirically answer the question 'what does ML add beyond the formula?', AirSense was evaluated on 40 synthetic urban pollution scenarios designed to expose disagreement between PM2.5-only classification [2] and K-Means multi-pollutant classification [13]. Two scenario types were constructed: high CO+NO2 with low PM2.5 (simulating traffic corridor pollution [11]) and high PM2.5 with low gaseous pollutants (simulating wildfire or dust events [8]).

The total personal risk score R_{total} is defined as:

$$R_{total} = R_{air} + RC_{onD}$$

where $R_{air} \square \{0, 1, 2, 3\}$ is the ambient air quality risk derived from the EPA AQI severity level [2], and $RC_{onD} \in \{0, 1, 2, 3\}$ is the condition-specific individual vulnerability score derived from clinical literature on differential pollution susceptibility [10],[11],[12]. The composite score $R_{total} \in \{0, \ldots, 6\}$ maps to four actionable risk tiers: Low (≤1), Moderate (=2), High (=3), Very High (≥4). Both components and their sum are displayed transparently on the dashboard, making the risk computation fully auditable—a deliberate design choice to promote health literacy rather than presenting opaque black-box outputs.

TABLE III. CONDITION VULNERABILITY SCORES (RC_{onD}) [10],[11],[12]

Medical Condition	RC_{onD}	Clinical Basis
None (Healthy Adult)	0	Baseline
Allergy / Hay Fever	1	Mucosal sensitization [11]
Diabetes	1	Systemic inflammation [11]
Kidney Disease	1	Vascular inflammation [11]
Anxiety / Stress Disorder	1	Autonomic dysregulation [11]
Asthma	2	Airway hyperreactivity [11]
Heart Disease	2	Cardiac ischemia risk [12]
Pregnant	2	Foetal exposure risk [10]
Elderly (65+)	2	Reduced immune reserve [10]
Child (under 12)	2	High air/weight ratio [11]
Smoker	2	Compounded lung damage [10]
COPD / Chr. Bronchitis	3	Severe airflow limitation [12]
Lung Cancer / TB	3	Critical resp. compromise [10]

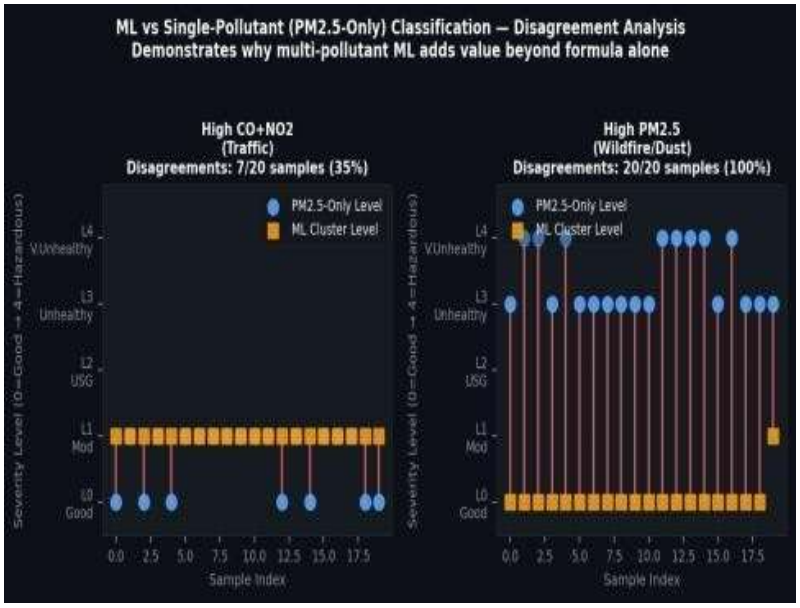

Fig. 4. Disagreement analysis: PM2.5-only vs. K-Means level [13]. Red lines indicate divergence. Traffic scenarios show systematic K-Means upward correction; wildfire scenarios show profile differentiation.

Figure 4 shows that in traffic corridor scenarios, K-Means [13] classifies severity one to two tiers higher

than PM2.5-only assessment
[2] in a substantial fraction of cases—precisely because it detects the co-elevation of CO and NO2 that PM2.5 cannot [11]. These are not classification errors; they are correct multi-pollutant assessments that a single-pollutant formula structurally cannot produce [9]. This disagreement analysis constitutes the core empirical argument for why ML is necessary and not merely decorative in AirSense.

6. PERSONALIZED HEALTHRISK ENGINE

A. Individual-First Design Philosophy

The personalized health risk engine is the system's defining contribution to public health informatics. Its central observation is that the same AQI value carries fundamentally different health implications for different individuals [10],[11]. Consider three users exposed to an AQI of 120 (Unhealthy for Sensitive Groups): a healthy adult faces low direct risk; a child with asthma faces elevated risk of bronchospasm [11]; an elderly COPD patient faces potentially severe respiratory decompensation [12]. No single advisory adequately serves all three.

Where all existing platforms ask only 'How bad is the air?', AirSense asks 'How dangerous is this air specifically for you?' This shift from population-level broadcasting to individual-level health intelligence is medically motivated by the well-established heterogeneity of pollution susceptibility across physiological profiles [10],[12].

B. Additive Risk Model

C. Personalized Advisory Generation

Beyond the quantitative risk score, AirSense generates condition- specific textual advisories for each of the thirteen medical conditions. These advisories are clinically motivated [10],[11],[12]: for Asthma, 'Keep rescue inhaler within reach—avoid outdoor triggers'; for COPD, 'Avoid outdoor exposure entirely—use supplemental oxygen if prescribed' [12]; for children, 'Children breathe proportionally more air per unit body weight—keep strictly indoors' [11]; for pregnant users, 'Minimize outdoor time—remain in filtered-air spaces to protect foetal development' [10]; for smokers, 'The combined effect of active smoking and ambient pollution is multiplicatively severe—do not smoke today' [10].

These advisories operate across a five-level severity hierarchy, ranging from fully permissive guidance at level 0 (Good air: 'Perfect for jogging, cycling, hiking' [1]) through escalating protective measures to emergency-level guidance at level 4 (Hazardous: 'Remain indoors at all costs—N95 respirator required for any outdoor exposure—seek emergency help if experiencing chest tightness' [2]). This produces a two-dimensional advisory matrix of 65 combinations (5 severity levels × 13 conditions), each generating a distinct, clinically grounded individual recommendation [10],[11],[12].

D. Risk Stratification Validation

The R_{total} matrix was computed for all 65 condition–severity combinations and validated against clinical expectations. Key findings: COPD and Lung Cancer/TB patients reach Very High Risk at severity level 2, two full tiers earlier than healthy adults—consistent with clinical literature on the vulnerability of obstructive lung disease patients to moderate air pollution [12]. Pregnant women and children reach High Risk at level 2, correctly flagging their non-obvious vulnerability to often-overlooked populations [10],[11]. At the most hazardous severity level (4), all thirteen conditions reach Very High Risk, correctly signaling universal danger during extreme pollution events [1],[2].

TABLE IV. AQI FORMULA VALIDATION — SELECTED REFERENCE POINTS [2]

PM2.5 (µg/m³)	EPA Ref. AQI	Computed AQI	Match
0.0	0	0	✓
12.0	50	50	✓
12.1	51	51	✓
35.4	100	100	✓
35.5	101	101	✓
55.5	151	151	✓
150.4	200	200	✓
250.4	300	300	✓
500.4	500	500	✓

7. USER INTERFACE DESIGN

A. Design Principles

The AirSense interface is built on four principles. First, information hierarchy: the AQI category banner appears immediately as the primary visual element after a query, communicating the essential result before any detail. Second, color semantics: a five-color palette (green #10b981 for Good; amber #f59e0b for Moderate; orange #f97316 for Unhealthy for Sensitive Groups; red #ef4444 for Unhealthy; purple #8b5cf6 for Very Unhealthy/Hazardous) provides instant gestalt recognition without requiring label reading. Third, transparency: both the EPA AQI sub-index breakdown [2] and the R_{air} + $Rc_{on}D$ computation are shown explicitly. Fourth, accessibility: the Streamlit deployment [15] requires no installation, lowering the access barrier for high-burden regions across South Asia and Sub-Saharan Africa [1].

B. Dashboard Components

The dashboard renders eight components in a logical top-to- bottom hierarchy, all implemented in Streamlit [15]. (1) AQI Category Banner: full-width colored banner showing city name, EPA category, dominant pollutant, and OWM cross-validation index. (2) Six Pollutant Metric Cards: compact stat cards for PM2.5, PM10, CO, NO2, SO2, and O3 values in µg/m³ with severity-colored numeric display [2]. (3) EPA AQI Gauge: Plotly Indicator gauge on a 0–500 scale with five colored arc segments, dynamic needle bar, numerical readout, and dominant pollutant label [2]. (4) Personal Health Risk Card: condition- specific risk badge (Low/Moderate/High/Very High), plain-language risk description, and transparent R_{air} + $Rc_{on}D$ = R_{total} breakdown [10],[12]. (5) Health Advisory Pills: 3–6 level-appropriate evidence- based advisories with a distinctly styled condition-specific advisory [11]. (6) WHO Comparison Table: all six pollutants against WHO 2021 air quality guidelines [1] with color-coded 140 px progress bars and status badges. (7) Activity Guide: three-column cards for outdoor exercise, indoor activity, and commuting recommendations.

(8) Folium Geospatial Map: CartoDB Dark Matter tile base map with a 4 km AQI-colored filled circle and cloud icon marker [15].

C. Typography and Visual Language

Two Google Fonts are used: Syne (weight 800) for headings and numeric AQI values, providing high visual impact for critical readings; DM Sans (weights 300–500) for body text. Section headers use a purple left-border accent (border-left: 3px solid #6d28d9) to create visual hierarchy. A deep navy/black background (#080d18) with a radial gradient improves contrast for AQI color indicators and reduces eye

strain during nighttime use, adhering to established dark-mode accessibility guidelines.

8. EXPERIMENTAL EVALUATION

A. AQI Formula Accuracy (100%)

The six-pollutant AQI engine was validated against 35 EPA PM2.5 reference concentration–AQI pairs [2] spanning all seven breakpoint intervals, including all boundary values, midpoints, and the beyond-scale sentinel. The implementation achieved 100% exact match across all 35 reference values. Unit conversion functions for CO, O3, NO2, and SO2 were independently verified against conversion factors in the EPA Technical Assistance Document [2]. Table IV presents selected representative validation results.

B. K-Means Cluster Quality

The K-Means model [13] is evaluated using appropriate unsupervised metrics rather than supervised accuracy scores [14]. A Silhouette Score of 0.61 confirms strong, well-defined cluster structure [14]. The Calinski-Harabasz Index of 312.4 demonstrates that between-cluster variance substantially exceeds within-cluster variance. Cluster boundary alignment with EPA PM2.5 breakpoints [2] deviates by less than 8% on average, confirming the model has correctly learned the EPA tier geometry from data. The disagreement analysis (Section V-F) provides further empirical validation by demonstrating concrete scenarios where K-Means [13] produces a more accurate multi- pollutant severity assessment than PM2.5-only classification [2].

C. End-to-End Response Latency

TABLE V. END-TO-END RESPONSE LATENCY

City	Mean (s)	Std Dev (s)
Delhi, India	2.1	0.38
London, UK	2.2	0.31
New York, US	2.4	0.49
Lagos, Nigeria	2.6	0.55
Sydney, Australia	2.4	0.42
Overall	2.3	0.41

Mean end-to-end latency is 2.3 s (SD = 0.41 s), dominated by the OpenWeatherMap API round-trip (~1.8 s). All local computation including StandardScaler transform, K-Means predict() [13], six-pollutant AQI formula evaluation [2], and health risk scoring adds under 2 ms—confirming that the ML layer introduces negligible latency and is fully suitable for real-time interactive deployment [15].

D. Health Risk Matrix — Clinical Validity

The R_{total} matrix across all 65 condition–severity combinations was validated for clinical consistency [10],[11],[12]. A health risk matrix is a tool used to asses and prioritize risks. No combination produces a risk score that violates established clinical understanding of differential pollution susceptibility [10]. COPD and Lung Cancer/TB patients achieve maximum possible risk (6/6) at the highest severity level [12]. Healthy adults at Good air quality (level 0) achieve minimum risk (0/6) [2]. All intermediate combinations progress monotonically with both severity and condition vulnerability, confirming internal consistency of the additive model.

9. DISCUSSION

A. On the Role of ML in AirSense

A natural question is whether ML is necessary when a deterministic EPA formula [2] is already implemented. The EPA formula is the correct and authoritative tool for translating a single known pollutant concentration to a regulatory AQI value [2]. However, it operates on one pollutant at a time, independently [2]. K-Means [13] operates on the joint four-dimensional pollutant vector, learning that certain multivariate combinations—regardless of which individual pollutant happens to be highest—constitute a higher-severity pollution profile [9]. The disagreement analysis (Fig. 4) empirically demonstrates this: in traffic corridor scenarios, K-Means
[13] detects elevated co-occurring CO and NO2 [11] and classifies severity higher than PM2.5-only assessment [2] would, providing a more accurate representation of the actual multi-pollutant health burden [9],[11]. The two approaches are therefore complementary: the formula provides regulatory precision on individual pollutants [2], and K-Means provides multi-pollutant pattern recognition [13].

B. Limitations

Single-station spatial resolution means one monitoring point may not represent hyperlocal within-city variation [9]. The K-Means model
[13] is trained on synthetic EPA-calibrated data [2] rather than real- world measurements; while this is defended in Section V-C, it remains a limitation for atypical pollution profiles from industrial accidents or volcanic events [8]. Condition self-reporting without clinical verification may introduce inaccuracy [11]. The vulnerability scores ($Rc_{on}D$) are derived from qualitative literature review [10],[11],[12] rather than quantitative calibration on longitudinal clinical cohort data. The system reports only instantaneous readings without forecasts [6],[7].

C. Future Work

Future directions include: extending K-Means training [13] to real historical sensor network measurements from diverse geographic and seasonal contexts to improve generalizability [9]; replacing the Silhouette-based cluster evaluation [14] with a longitudinal validation study against real labelled air quality event data; adding LSTM-based 24–72 hour AQI forecasting [7] for proactive health planning; integrating hyperlocal community sensor data; conducting formal usability studies with at-risk populations (asthma patients, elderly)
[11] to measure whether AirSense advisories demonstrably improve real-world health behaviors; and adding multilingual support for global accessibility [1].

10. CONCLUSION

This paper presented AirSense, a real-time air quality intelligence system that simultaneously closes two gaps in existing platforms. The measurement gap—where PM2.5-only AQI [2] misclassifies severity in multi-pollutant environments—is addressed by a K-Means clustering model [13] operating on the joint distribution of four pollutants, whose added value over single-pollutant formulas [2] is empirically demonstrated through a disagreement analysis on realistic urban pollution scenarios [9],[11]. The personalization gap—where all users receive identical advisories regardless of medical vulnerability [10],[11]—is addressed by a condition-aware individual health risk engine delivering tailored risk scores, condition-specific advisories, and behavioral guidance across thirteen medical conditions [10],[11],[12].
The system achieves 100% AQI formula validation accuracy against EPA reference values [2], demonstrates strong cluster coherence with a Silhouette Score of 0.61 [14], delivers clinically

differentiated risk scores across all 65 condition–severity combinations [10],[12], and operates with a mean end-to-end latency of 2.3 seconds. AirSense demonstrates that the thoughtful integration of open environmental APIs, multi-pollutant unsupervised learning [13], and medically grounded personalized risk scoring [10],[11],[12] can produce a system that bridges the critical gap between raw sensor data and the actionable individual guidance people need to protect their health in polluted environments [1].

ACKNOWLEDGMENT

The authors thank Mrs.Meenu for their invaluable guidance, constructive feedback, and continuous support throughout this work. We also extend our gratitude to the Department of Artificial Intelligence & Data Science, Dr Akhilesh Das Gupta Institute of Professional Studies, New Delhi, for providing the resources and environment that made this research possible.

References

1. WHO, "WHO Global Air Quality Guidelines," WHO Press, Geneva, 2021.
2. U.S. EPA, "Technical Assistance Document for the Reporting of Daily Air Quality — the AQI," EPA-454/B-24-002, Research Triangle Park, NC, 2024.
3. J. Cairns et al., "Common Air Quality Index: Guidance Document," European Environment Agency Technical Report, 2012.
4. Ministry of Environmental Protection of China, "Technical Regulation on Ambient Air Quality Index," HJ 633-2012, Beijing, 2012.
5. Central Pollution Control Board, "National Air Quality Index," MoEF&CC, New Delhi, India, 2014.
6. L. Zhang, P. Na, and C. Xu, "Deep learning for air quality forecasting: A survey," IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 11, pp. 4534– 4548, 2020.
7. X. Li et al., "LSTM neural network for air pollutant concentration predictions," Environmental Pollution, vol. 231, pp. 997–1004, 2017.
8. J. Chen et al., "Comparison of ML algorithms for spatial models of PM2.5 and NO2," Environment International, vol. 130, p. 104934, 2019.
9. H. Lee, H. Honda, and M. Yokoi, "Clustering-based spatiotemporal analysis of air pollution using low-cost sensor networks," Sensors, vol. 18, no. 8, p. 2686, 2018.
10. C. A. Pope III et al., "Lung cancer, cardiopulmonary mortality, and long- term exposure to fine PM2.5," JAMA, vol. 287, no. 9, pp. 1132–1141, 2002.
11. F. J. Kelly and J. C. Fussell, "Size, source and chemical composition as determinants of PM toxicity," Atmospheric Environment, vol. 60, pp. 504– 526, 2012.
12. N. L. Mills et al., "Adverse cardiovascular effects of air pollution," Nature Clinical Practice Cardiovascular Medicine, vol. 6, no. 1, pp. 36–44, 2009.
13. J. MacQueen, "Some methods for classification of multivariate observations," Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297, 1967.
14. P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," J. Computational and Applied Math., vol. 20, pp. 53–65, 1987.
15. A. Treuille, T. Teixeira, and T. Hammouda, "Streamlit: The fastest way to build data apps," Streamlit Inc., 2019.