Aerocare: A Regulation-Faithful Six-Pollutant Air Quality Intelligence System

with Unsupervised Severity Profiling and Condition-Aware Health Risk Scoring

***Lakshay Giri*[1]**
Student, B.Tech(AI&DS)

***Shrut Jain*[2]**
Student, B.Tech(AI&DS)

***Nimish Gupta*[3]**
Student, B.Tech(AI&DS)

***Mr. Ritesh Kumar*[4]**
Assistant Professor

Department of Artificial Intelligence & Data Science Dr. Akhilesh Das Gupta Institute of Professional Studies New Delhi, India

Abstract—Air quality platforms are only as trustworthy as the regulatory arithmetic beneath them. This paper presents Aerocare, the second-generation implementation of the AirSense air quality intelligence system [16], re-engineered around three changes. First, the AQI engine was rebuilt as a regulation-faithful six-pollutant calculator aligned with the 2024 US EPA breakpoints [2], including pollutant-specific truncation rules, unit conversion of OpenWeatherMap concentrations, and averaging-period logic for SO2 (1-hour/24-hour) and O3 (8-hour/1-hour). The engine reproduces all 78 tabulated breakpoint endpoints exactly, and a head-to-head comparison shows that the 2024 PM2.5 revision alone moves 25% of the 100–200 µg/m³ range into a higher severity level. Second, the unsupervised K-Means severity model [13] was extended from four to six pollutants and re-evaluated with a diagnostic protocol (silhouette 0.48, Calinski–Harabasz 245.0, Davies–Bouldin 0.84, adjusted Rand index 0.70 against generating tiers). The three-way comparison against PM2.5-only and six-pollutant EPA classification shows that, once every sub-index enters the EPA maximum, the formula already resolves the PM2.5-only blind spot, and the learned clusters are best used as a pollution-signature profile rather than a replacement severity classifier. Third, the condition-aware additive health risk engine [10]–[12] was enumerated across thirteen conditions and five severity levels (65 combinations: 11 Low, 17 Moderate, 15 High, 22 Very High), exposing a baseline-vulnerability floor not previously reported. Local computation adds 1.3 ms per request. We argue that reporting where machine learning does and does not help is as useful to practitioners as a headline accuracy figure.

Keywords—*air quality monitoring, US EPA AQI 2024 revision, six-pollutant AQI, K-Means clustering, cluster validity, personalized health risk, eight-hour ozone average, Streamlit, environmental informatics*

I. INTRODUCTION

Ambient air pollution causes approximately 4.2 million premature deaths per year globally, predominantly through stroke, heart disease, chronic obstructive pulmonary disease (COPD), and lung cancer [1]. Six criteria pollutants are of 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].

In earlier work we introduced AirSense [16], a real-time system that closed two gaps in consumer air quality platforms: a measurement gap (indices computed from PM2.5 alone) and a personalization gap (a single broadcast value for all users). Continued development and a careful re-reading of that implementation exposed three further issues that motivate the present paper. First, regulatory drift: the US EPA revised its PM2.5 breakpoints in the 2024 Technical Assistance Document [2], and a system that reports an official index must follow the current specification. Second, incomplete pollutant coverage: the earlier validation exercised only the PM2.5 sub-index, while the engine computes six, and the averaging-period rules that govern SO2 and O3 were not modelled. Third, an under-specified role for machine learning: the K-Means layer used a four-pollutant feature vector, and its evaluation did not separate what the six-pollutant maximum rule of the EPA formula already provides from what a joint-distribution model might add.

This paper presents Aerocare, the redesigned system, and reports its evaluation candidly, including results that do not favour the machine-learning component. The contributions are:

- A six-pollutant US EPA AQI engine aligned with the 2024 breakpoints [2], with pollutant-specific truncation, unit conversion, and SO2/O3 averaging-period logic, verified on 78 tabulated breakpoint endpoints.
- A quantitative comparison of legacy and 2024 PM2.5 breakpoints showing exactly which concentration ranges change severity level.
- A standalone eight-hour ozone averaging module using the Open-Meteo Air Quality API [21] with an eight-of-eight completeness rule.
- A six-feature K-Means severity profiler [13], evaluated with cluster validity indices [14],[18]–[20] and a three-way comparison against PM2.5-only and six-pollutant EPA classification.
- A fully enumerated condition-aware additive risk matrix (13 conditions × 5 severity levels) with an analysis of its tier structure and its baseline floor.
- A refined Streamlit dashboard [15] with eight components, a WHO 2021 comparison [1], an activity guide, and a geospatial map.

II. RELATED WORK

A. AQI Standards and the 2024 PM2.5 Revision

The US EPA AQI [2] maps each pollutant concentration to a 0–500 index through piecewise linear interpolation, and the composite AQI is the maximum sub-index across pollutants. The 2024 Technical Assistance Document tightens the PM2.5 breakpoints: the Good category ends at 9.0 µg/m³ and the upper categories are redefined (Unhealthy 55.5–125.4, Very Unhealthy 125.5–225.4, Hazardous 225.5

and above) [2]. International equivalents include the European CAQI [3], China's AQI [4], and India's NAQI [5]. Most consumer platforms still implement only the PM2.5 sub-index [8].

B. Machine Learning for Air Quality

Random Forest, Support Vector Regression, and LSTM models have been applied extensively to AQI forecasting [6],[7]. Classification approaches have categorized severity from multivariate sensor data [8], and unsupervised methods such as K-Means and DBSCAN have identified spatial pollution clusters in monitoring networks [9]. Evaluating unsupervised models requires internal validity indices, such as the silhouette coefficient [14], the Calinski–Harabasz index [18], and the Davies–Bouldin index [19], and, where reference labels exist, partition-agreement measures such as the adjusted Rand index [20].

C. Personalized Exposure Guidance

Epidemiological evidence documents differential susceptibility to air pollution across health states [10],[11],[12]. Platforms such as AirNow, OpenAQ, AirVisual, and IQAir offer at most a binary "sensitive groups" flag [11]. AirSense [16] introduced an additive risk model combining ambient severity with a condition-specific vulnerability score; Aerocare retains this model and analyses its behaviour exhaustively.

III. SYSTEM ARCHITECTURE

A. Four-Layer Pipeline

Aerocare is a four-layer pipeline: (1) Data Acquisition from the OpenWeather Geocoding and Air Pollution APIs [22], plus a standalone Open-Meteo ozone module [21]; (2) Processing, in which the deterministic six-pollutant EPA AQI computation [2] and the K-Means inference [13] execute on the same observation; (3) Personalized Health Risk Computation, which combines the EPA-derived severity level with the user's declared condition [10],[12]; and (4) Visualization through eight dashboard components in Streamlit [15]. In the current build the user-facing banner, gauge, risk score, and advisories are driven by the EPA-derived level; the K-Means level is computed for each request and persisted in the result record for diagnostic comparison.

B. Technology Stack

The system is written in Python and uses Streamlit [15] for the interface, Pandas and NumPy for data handling, scikit-learn [17] (StandardScaler, KMeans) for the ML layer, Plotly for the AQI gauge, Folium with streamlit-folium for mapping, and the Requests library for HTTP access. Custom CSS with Google Fonts (Syne, DM Sans) implements a dark navy theme (#080d18).

C. Data Acquisition, Session State and Error Handling

A city name is geocoded through the OpenWeather Geocoding API, and the returned coordinates are used to query the Air Pollution API, which provides all six concentrations in µg/m³ together with OpenWeather's own five-point index. Both requests enforce a 10-second timeout. Results are stored in Streamlit session state, so interacting with widgets after the initial fetch does not trigger further API calls, and a schema guard discards any stored result that lacks the keys required by the current version. Error handling is differentiated: geocoding returns an explicit "location not found" message for empty results; HTTP 401 produces an invalid-credential message; other HTTP errors, connection failures, empty pollutant lists, and malformed or incomplete JSON each produce a distinct, actionable message rather than a generic failure notice.

D. Eight-Hour Ozone Module

The EPA O3 8-hour sub-index requires an eight-hour average [2]. The function get_o3_8h_average queries the Open-Meteo Air Quality API [21] for hourly ozone (past day plus forecast day, UTC), selects the eight hourly values in the window [t−8 h, t) where t is the current hour, and returns their mean only if all eight values are present; otherwise it returns no value. Its request timeout is 15 seconds. Section 9.B describes the current integration status of this module.

IV. AQI COMPUTATION METHODOLOGY

A. Piecewise Linear Formula

Each sub-index is computed with the EPA interpolation [2]:

$$I = [(I_{hi} - I_{lo}) / (C_{hi} - C_{lo})] \times (C_p - C_{lo}) + I_{lo}$$

where C_p is the truncated concentration and C_{lo}, C_{hi}, I_{lo}, I_{hi} are the bounding concentrations and index values of the interval containing it. The sub-index is rounded to the nearest integer. Concentrations above the highest breakpoint return the sentinel value 500. The composite AQI is the maximum of the available sub-indices, and the pollutant producing it is reported as the dominant pollutant; if no sub-index is available the system reports "Unavailable" rather than a numeric value.

B. Truncation and Unit Conversion

EPA specifications require each concentration to be truncated, not rounded, before interpolation [2]. Aerocare applies: PM2.5 to one decimal (µg/m³); PM10 to an integer (µg/m³); CO to one decimal (ppm); NO2 to an integer (ppb); SO2 to an integer (ppb); and O3 to three decimals (ppm). OpenWeather reports all pollutants in µg/m³, so the following conversions precede truncation: CO ÷ 1,145.4 (→ ppm); NO2 ÷ 1.88 (→ ppb); SO2 ÷ 2.62 (→ ppb); O3 ÷ 1,961 (→ ppm). PM2.5 and PM10 are used directly [2].

C. Averaging-Period Logic for SO2 and O3

Two pollutants have breakpoint tables defined on different averaging periods [2]. For SO2, the 1-hour value is used up to AQI 200 (304 ppb); when the 1-hour value is 305 ppb or higher, the 24-hour value must be used for AQI above 200. Aerocare implements this branch and, when a 24-hour value is not supplied, caps the sub-index at 200 rather than extrapolating. For O3, the 8-hour table covers AQI 0–300 and the 1-hour table covers AQI 101–500; when both averages are supplied the larger sub-index is reported, and the 8-hour table returns the sentinel 500 above 0.200 ppm.

D. Implemented Breakpoints

TABLE I. PM2.5 BREAKPOINTS: 2024 REVISION (IMPLEMENTED) VS. LEGACY [2]

2024 C_{lo}	2024 C_{hi}	AQI range	Category	Legacy C_{hi} (µg/m³)
0.0	9.0	0–50	Good	12.0
9.1	35.4	51–100	Moderate	35.4
35.5	55.4	101–150	Unhl. (Sensitive)	55.4
55.5	125.4	151–200	Unhealthy	150.4
125.5	225.4	201–300	Very Unhealthy	250.4
225.5	325.4	301–500	Hazardous	350.4 / 500.4*

*Legacy tables split the 301–500 range into two intervals (301–400, 401–500).

TABLE II. IMPLEMENTED BREAKPOINTS FOR THE REMAINING POLLUTANTS [2]

Category	PM10 (µg/m³)	CO (ppm)	NO2 (ppb)	SO2 (ppb)	O3 8-h (ppm)
Good (0–50)	0–54	0.0–4.4	0–53	0–35	0.000–0.054
Moderate (51–100)	55–154	4.5–9.4	54–100	36–75	0.055–0.070
USG (101–150)	155–254	9.5–12.4	101–360	76–185	0.071–0.085
Unhealthy (151–200)	255–354	12.5–15.4	361–649	186–304	0.086–0.105
Very Unhealthy (201–300)	355–424	15.5–30.4	650–1249	305–604†	0.106–0.200
Hazardous (301–500)	425–604	30.5–50.4	1250–2049	605–1004†	(1-h O3: 0.405–0.604)

†24-hour SO2 averages. O3 1-hour table: 0.125–0.164 (101–150), 0.165–0.204 (151–200), 0.205–0.404 (201–300), 0.405–0.604 (301–500).

V. MACHINE LEARNING MODULE

A. Role of the Learned Model

The EPA composite AQI takes the maximum of six independent single-pollutant assessments [2]; it does not model the joint multivariate relationship between pollutants. Different pollution sources produce characteristic multi-pollutant signatures: traffic corridors elevate CO and NO2 together, industrial events elevate SO2 and NO2, and biomass or dust episodes elevate particulates with suppressed gases [8],[9],[11]. K-Means [13] operating on the joint feature vector can in principle summarize such signatures. In Aerocare the question is posed as an empirical one: given that the maximum rule now covers all six pollutants, what does the learned model add? Section 8.C answers it.

B. Feature Vector and Training Data

The feature vector was extended from four to six pollutants: [PM2.5, PM10, CO, NO2, SO2, O3], all in µg/m³. Training data are synthetic: for each of five severity levels, 20 samples are drawn uniformly at random within a per-level concentration range for every pollutant (Table III), giving 100 samples (NumPy seed 42). Synthetic generation guarantees reproducibility and full coverage of the severity ranges, but it also means the model learns the geometry of the generator rather than of real urban atmospheres; Section 9.B discusses the consequences.

TABLE III. TRAINING-DATA GENERATOR RANGES (µg/m³, 20 SAMPLES PER LEVEL)

Level	PM2.5	PM10	CO	NO2	SO2	O3
L0 Good	0–12	0–54	0–4400	0–53	0–35	0–54
L1 Moderate	12–35	54–154	4400–9400	53–100	35–75	54–70
L2 USG	35–55	154–254	9400–12400	100–360	75–185	70–85
L3 Unhealthy	55–150	254–354	12400–15400	360–649	185–304	85–105
L4 V.Unh./Haz.	150–500	354–604	15400–30400	649–1249	304–1004	105–200

C. Preprocessing and Cluster Ordering

All six features are standardized with StandardScaler (zero mean, unit variance) before clustering, which is essential because CO spans tens of thousands of µg/m³ while O3 spans hundreds; without standardization CO would dominate the Euclidean distance [13]. K-Means is run with k = 5, n_init = 20, and random_state = 42. Centroids are inverse-transformed to concentration space and ranked by PM2.5 centroid value, which defines a mapping from cluster identifier to severity level {0,…,4}.

D. Selecting k

Within-cluster inertia for k = 2,…,9 is 186.2, 91.6, 72.0, 53.0, 41.9, 35.5, 31.4, and 28.6 (Fig. 1). The curve falls steeply to k = 3 and then declines gradually; there is no sharp elbow at k = 5. The mean silhouette score is highest at k = 2 (0.659) and is 0.481 at k = 5, with a secondary maximum of 0.519 at k = 6. We therefore state plainly that k = 5 is chosen a priori to match the five severity levels exposed by the interface, and is not discovered from the data.

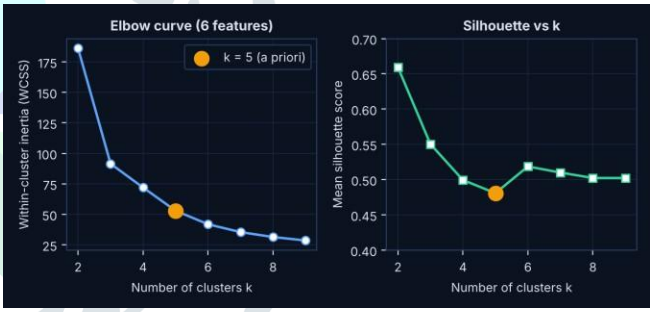

Fig. 1. Elbow curve and silhouette score for k = 2 to 9 on the six-feature standardized training set. k = 5 (marked) is set by the interface's five severity levels.

E. Cluster Quality

Clustering is unsupervised, so supervised "accuracy" is not the appropriate summary for real data [14]. Table IV reports internal validity indices on the standardized training set and, because the generating tier of every synthetic sample is known, two partition-agreement indices [20].

TABLE IV. CLUSTER QUALITY (SIX FEATURES, k = 5, STANDARDIZED SPACE)

Metric	Value	Interpretation
Mean distance to own centroid	0.64	Compact clusters
Mean inter-centroid distance	3.95	Centroids well separated
Silhouette score [14]	0.481	Reasonable structure (below k = 2 and k = 6)
Calinski–Harabasz index [18]	245.0	Between-cluster variance dominates
Davies–Bouldin index [19] (lower is better)	0.844	Moderate overlap between neighbours
Adjusted Rand index vs. generating tier [20]	0.697	Substantial but imperfect agreement
Normalized mutual information vs. tier	0.842	High shared information

The first two principal components of the standardized data explain 89.8% and 4.1% of the variance (Fig. 2). Fig. 3 shows the centroid profiles in concentration space. Two features of these profiles matter for interpretation. First, ordering clusters by PM2.5 alone does not yield monotonic profiles in the gaseous pollutants: the L4 centroid has the highest PM2.5 (421 µg/m³) but lower NO2, SO2, and O3 than the L3 centroid, i.e., the two upper clusters differ in *composition* as much as in severity. Second, the clusters do not coincide with the generating tiers (Table V): the Moderate and USG tiers fall almost entirely into a single cluster (39 samples), while the highest tier is split between two clusters (11 and 9 samples).

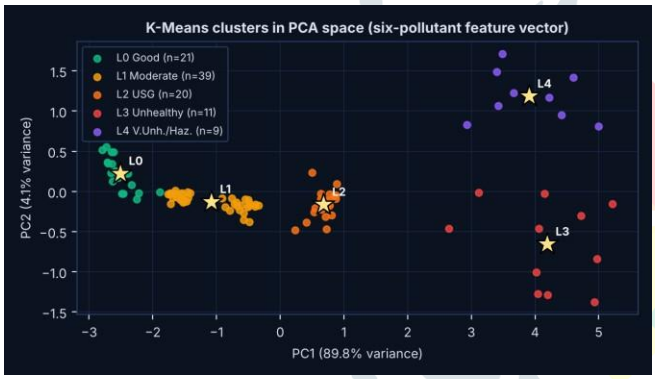

Fig. 2. K-Means clusters projected to two principal components. Stars mark centroids, labelled with the severity level assigned by PM2.5 ranking.

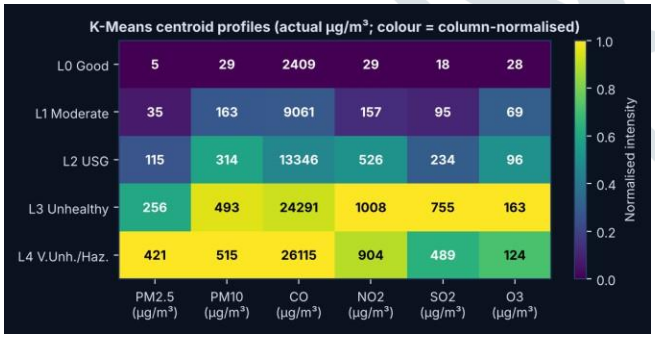

Fig. 3. Cluster centroid heatmap (actual µg/m³ values). Colour is normalised per pollutant.

TABLE V. GENERATING TIER VS. ASSIGNED ML LEVEL (TRAINING SET, n = 100)

Generating tier	ML L0	ML L1	ML L2	ML L3	ML L4
L0 Good	20	0	0	0	0
L1 Moderate	1	19	0	0	0
L2 USG	0	20	0	0	0
L3 Unhealthy	0	0	20	0	0
L4 V.Unh./Haz.	0	0	0	11	9

VI. PERSONALIZED HEALTH RISK ENGINE

A. Additive Risk Model

The central observation of the engine is that the same AQI carries different implications for different individuals [10],[11]. The total personal risk score is

$$R_{\text{total}} = R_{\text{air}} + R_{\text{cond}}$$

where $R_{air} \in \{0,1,2,3\}$ is derived from the EPA severity level and $R_{cond} \in \{0,1,2,3\}$ is a condition-specific vulnerability score drawn from the clinical literature [10],[11],[12] (Table VI). The mapping from severity level (0–4) to R_{air} is [0, 0, 1, 2, 3]: Good and Moderate air add no ambient risk, USG adds 1, Unhealthy adds 2, and Very Unhealthy/Hazardous adds 3. The sum $R_{total} \in \{0,\ldots,6\}$ maps to four tiers: Low (≤ 1), Moderate (= 2), High (= 3), and Very High (≥ 4). Both components and their sum are displayed on the dashboard so that the computation is auditable.

TABLE VI. CONDITION VULNERABILITY SCORES (R_{cond}) [10],[11],[12]

Medical condition	R_{cond}	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 respiratory compromise [10]

B. Advisory Generation

Advisories are composed at run time from two independent sources. A level advisory set contains three to six pills per severity level (3, 3, 4, 5, and 6 pills for levels 0–4), ranging from permissive guidance in Good air ("Perfect for jogging, cycling, hiking") to emergency guidance at level 4 (remain indoors, N95 respirator for any outdoor exposure, medication immediately accessible, call emergency services for severe symptoms). A condition advisory is added for each of the twelve non-baseline conditions, for example "Keep rescue inhaler within reach" for asthma, "Do NOT smoke today" for smokers, and "Children breathe more air per body weight" for children. The condition text is independent of the level, so the 65 level–condition combinations yield 65 distinct composed advisories, of which the condition-specific part is constant for a given condition. A three-column activity guide (outdoor exercise, indoor activity, commuting) is also level-driven. All advisory text is informational and is not medical advice.

VII. USER INTERFACE DESIGN

A. Design Principles

The interface follows four principles: information hierarchy (the AQI category banner appears first); colour semantics (green #10b981 Good, amber #f59e0b Moderate, orange #f97316 USG, red #ef4444 Unhealthy, purple #8b5cf6 Very Unhealthy/Hazardous); transparency (the sub-index of the dominant pollutant and the $R_{air} + R_{cond} = R_{total}$ breakdown are shown explicitly); and accessibility (browser-based deployment with no installation, lowering the barrier in high-burden regions [1]).

B. Dashboard Components

The dashboard renders eight components top to bottom. (1) AQI category banner with city, EPA category, and the OpenWeather 1–5 index as a cross-reference. (2) Six pollutant cards in µg/m³. (3) EPA AQI gauge: a Plotly indicator on a 0–500 scale with five arc segments (0–50, 50–100, 100–200, 200–300, 300–500), whose title shows the category, the dominant pollutant, and that pollutant's sub-index. (4) Personal health risk card with the tier badge, a plain-language description, and the additive breakdown. (5) Health advisory pills with a distinctly styled condition advisory. (6) Activity guide. (7) WHO comparison table using the 2021 guideline values [1] (PM2.5 15, PM10 45, CO 4,000, NO2 25, SO2 40, O3 100 µg/m³), with a 140 px bar capped at 100% and a status of Safe (below 1× the guideline), Elevated (1–2×), or Dangerous (2× or more). (8) Folium map on a CartoDB Dark Matter base with a 4 km severity-coloured circle and a marker that is red at level 3 or above and green otherwise.

VIII. EXPERIMENTAL EVALUATION

All experiments call the functions and model-training code of the released application directly. Concentration data for the ML and scenario experiments are synthetic, because live-API evaluation against reference-grade monitors was outside the scope of this study; accordingly, no claim is made about real-world classification accuracy.

A. AQI Engine Validation

For every pollutant, the lower and upper endpoint of every breakpoint interval was converted to µg/m³ where required (using the factors of Section 4.B), passed through the corresponding sub-index function, and compared with the tabulated index value [2]. The engine matched 78 of 78 endpoints (100%), including the SO2 24-hour branch and both O3 averaging tables (Table VII). A dense scan of all 3,254 one-decimal PM2.5 values from 0.0 to 325.3 µg/m³ against an independent integer-arithmetic implementation found no mismatch. Edge cases behave as specified: PM2.5 above 325.4 µg/m³ returns 500; SO2 at or above 305 ppb without a 24-hour value returns 200; and an SO2 24-hour value of 400 ppb returns 232.

TABLE VII. BREAKPOINT-ENDPOINT VALIDATION BY POLLUTANT

Pollutant / table	Endpoints tested	Matches	Exact match
PM2.5	12	12	100%
PM10	12	12	100%
CO	12	12	100%
NO2	12	12	100%
SO2 (1-h and 24-h tables)	12	12	100%
O3 (8-hour table)	10	10	100%
O3 (1-hour table)	8	8	100%
Total	78	78	100%

One implementation detail deserves mention. CO is converted from µg/m³ to ppm by division before truncation, and binary floating point makes seven of 505 exactly representable tenth-of-ppm values (for example 7.5 and 15.0 ppm) truncate one tenth low. None of these crosses a category boundary, and measured concentrations are essentially never exact, so the practical effect is negligible; decimal arithmetic would remove it.

B. Impact of the 2024 PM2.5 Revision

Table VIII and Fig. 4 compare the legacy PM2.5 breakpoints with the 2024 breakpoints implemented in Aerocare. At the five-level resolution used by the interface, 5.6% of the PM2.5 range from 0 to 500.4 µg/m³ (scanned at 0.1 µg/m³) changes severity level. The change is concentrated in two ranges: 6.0% of the 0–50 µg/m³ range (concentrations of 9.1–12.0 µg/m³ move from Good to Moderate) and 25% of the 100–200 µg/m³ range (concentrations of 125.5–150.4 µg/m³ move from Unhealthy to Very Unhealthy). The magnitude of the index shift grows with concentration, for instance from 200 to 226 AQI at 150.4 µg/m³ and from 400 to 500 at 350.4 µg/m³. A system still using the legacy table would under-report severity precisely where public-health action is most needed.

TABLE VIII. PM2.5 SUB-INDEX UNDER LEGACY AND 2024 BREAKPOINTS

PM2.5 (µg/m³)	Legacy AQI	2024 AQI	Difference
9.0	38	50	+12
12.0	50	56	+6
35.4	100	100	0
55.5	151	151	0
100.0	174	182	+8
125.4	187	200	+13
150.4	200	226	+26
200.0	250	275	+25
250.4	300	351	+51
300.0	350	449	+99
350.4	400	500	+100

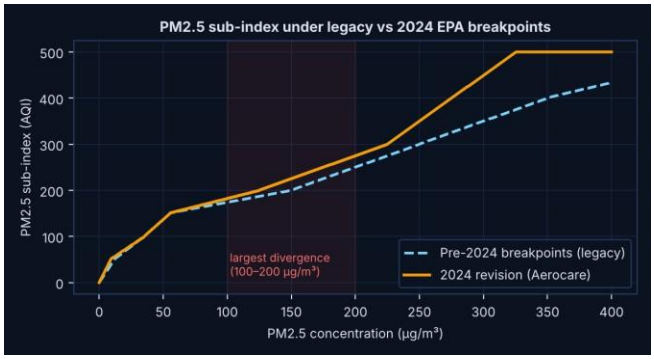

Fig. 4. PM2.5 sub-index under legacy and 2024 EPA breakpoints. The shaded band marks the 100–200 µg/m³ range with the largest change in severity level.

C. Comparative Classification Analysis

Three classifiers were compared on identical observations: (i) the PM2.5-only EPA level; (ii) the six-pollutant EPA level (maximum sub-index, Section 4); and (iii) the six-feature K-Means level (Section 5). Two experiments were run. In the first, a held-out synthetic set of 500 observations (100 per tier, same generator as Table III, different seed) was classified. In the second, two scenario families of 20 observations each were constructed to probe the cases that motivate the learned model: a *traffic corridor* (PM2.5 5–30, PM10 20–90, CO 12,400–25,000, NO2 360–900, SO2 2–30, O3 10–50 µg/m³) and a *smoke/dust episode* (PM2.5 55–300, PM10 150–450, CO 300–2,500, NO2 5–40, SO2 1–20, O3 10–60 µg/m³).

TABLE IX. AGREEMENT BETWEEN CLASSIFIERS ON THE HELD-OUT SET (n = 500)

Comparison	Exact agreement	Note
Six-pollutant EPA vs. PM2.5-only EPA	99.6%	Dominant pollutant mostly PM
Six-pollutant EPA vs. generating tier	91.0%	Legacy generator ranges
K-Means vs. generating tier	47.8%	Tier merging (Table V)
K-Means vs. six-pollutant EPA	43.2%	Weighted κ = 0.81
K-Means vs. PM2.5-only EPA	43.6%	—

The held-out result is sobering for the severity-classification use of K-Means. Exact agreement with the six-pollutant EPA level is 43.2%, and every disagreement is in the same direction: the learned level is lower than the EPA level (24 observations at EPA level 1, 100 at level 2, 78 at level 3, and 82 at level 4, usually by one level and in 22 cases by two). The quadratic-weighted kappa of 0.81 shows that the ordering is largely preserved, but the cluster-to-level mapping is shifted, a direct consequence of the tier merging and splitting seen in Table V. In the held-out EPA classification the dominant pollutant was PM2.5 in 271 observations, PM10 in 175, CO in 45, NO2 in 8, and O3 in 1; SO2 never dominated, a property of the generator ranges.

TABLE X. SCENARIO-FAMILY RESULTS (20 OBSERVATIONS PER FAMILY, MEAN SEVERITY LEVEL 0–4)

Scenario family	PM2.5-only	Six-pollutant EPA	K-Means	Key counts
Traffic corridor	0.75	3.60	1.05	PM2.5-only ≠ EPA: 20/20; K-Means > PM2.5-only: 6/20; K-Means < EPA: 20/20
Smoke / dust	3.65	3.70	0.75	PM2.5-only ≠ EPA: 1/20; K-Means < PM2.5-only: 20/20; K-Means < EPA: 20/20

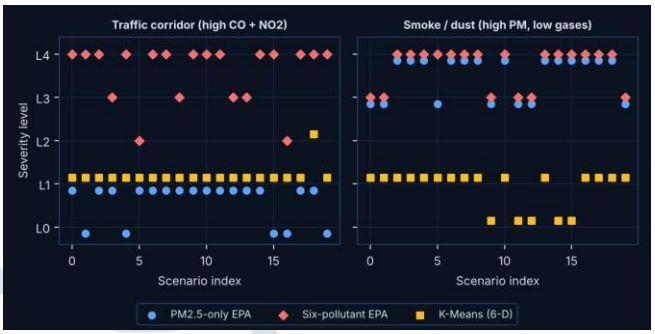

Fig. 5. Severity level assigned to each scenario by PM2.5-only EPA, six-pollutant EPA, and K-Means. In traffic scenarios the PM2.5-only level is far below the six-pollutant level.

The scenario results separate two effects that the earlier paper [16] conflated. The traffic scenarios confirm the *measurement-gap* argument: a PM2.5-only index classifies all 20 observations at Good or Moderate (mean level 0.75), whereas the six-pollutant maximum rule classifies them at a mean level of 3.60 because CO and NO2 sub-indices dominate. That gap is closed by the *formula*, not by machine learning. K-Means raises the level above the PM2.5-only value in 6 of 20 traffic observations, but remains below the six-pollutant EPA level in all 20, and in the smoke scenarios it lies below both formula-based levels in all 20 observations. On this evidence the learned model should not replace the formula for severity; its value is as a descriptive profile of pollutant composition, as illustrated by the compositional difference between the L3 and L4 centroids (Fig. 3).

D. Health Risk Matrix

The risk score was computed for all 65 combinations of condition and severity level (Fig. 6). The matrix is non-decreasing in both severity and condition vulnerability. Across the 65 cells, 11 fall in the Low tier, 17 in Moderate, 15 in High, and 22 in Very High. COPD and Lung Cancer/TB reach Very High at level 2, two levels earlier than healthy adults reach High. Asthma, heart disease, pregnancy, the elderly, children, and smokers reach High at level 2 and Very High at level 3. Allergy, diabetes, kidney disease, and anxiety reach High at level 3 and Very High at level 4. At the most severe level, 12 of 13 profiles are Very High; the healthy adult reaches High (R_{total} = 3) but not Very High. Two properties of the additive structure are important. First, because R_{cond} is added regardless of air quality, COPD and Lung Cancer/TB patients (R_{cond} = 3) are already in the High tier at level 0 (Good air), so the dashboard displays "High risk" for them on a clean-air day. Second, because levels 0 and 1 share R_{air} = 0, Good and Moderate air are not distinguished by the risk score.

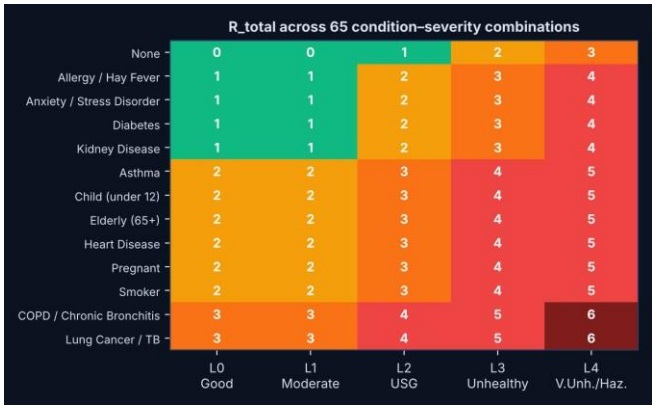

Fig. 6. Total risk score R_{total} for each condition and severity level. Green: Low; amber: Moderate; orange: High (= 3); red: Very High (≥ 4).

E. Computational Cost

Local computation per request (six-pollutant AQI, feature scaling, K-Means prediction, and level mapping) averaged 1.27 ms over 500 repeated calls on a single CPU process, of which the AQI engine alone accounts for about 5 µs. Fitting the scaler and the K-Means model on the 100 training samples takes about 12 ms; because Streamlit re-executes the script on each interaction, this cost recurs but remains negligible. End-to-end latency is dominated by the network round-trip to the data provider; the request path to OpenWeather is unchanged from the earlier system, whose measured mean end-to-end latency was 2.3 s [16]. Network latency was not re-benchmarked for this version.

IX. DISCUSSION

A. The Role of Machine Learning Revisited

The central empirical finding of this study is that regulatory fidelity delivered more than learning did. Completing the six-pollutant maximum rule removed the PM2.5-only blind spot entirely in the traffic scenarios, and correcting the PM2.5 breakpoints changed the severity level of one quarter of the 100–200 µg/m³ range. By contrast, the six-feature K-Means model, trained on synthetic data, reproduces the generator's overlaps and under-classifies severity relative to the formula. We conclude that, in this architecture, the formula should remain the sole authority for severity and risk, and the clustering layer should be positioned as a composition-profiling tool, a role in which the centroid analysis already provides interpretable signatures. Reporting this negative result avoids overstating the benefit of machine learning in public-health informatics.

B. Limitations

(1) *Synthetic training data.* The 100 training samples draw each pollutant independently within tier ranges, so they carry no real correlation structure, and the generator's PM2.5 boundaries (12, 35, 55, 150 µg/m³) retain the legacy values, whereas the AQI engine uses the 2024 table. As a result 9–10% of synthetic observations carry a generating label that differs from the six-pollutant EPA level. No evaluation against real, labelled monitoring data was performed. (2) *ML output not surfaced.* The K-Means level is computed and stored with each result but is not displayed; the runtime disagreement view described for the earlier system [16] is not rendered in this build. (3) *Averaging periods in the live path.* The dashboard currently supplies the instantaneous OpenWeather ozone concentration to the 8-hour O3 branch and supplies no 24-hour SO2 value, so the O3 sub-index is an approximation and SO2 at or above 305 ppb is capped at 200. The Open-Meteo eight-hour module (Section 3.D) is implemented but not yet connected to this path. (4) *Risk-model structure.* The additive model has a baseline floor (Section 8.D), treats Good and Moderate air identically, and uses qualitative vulnerability scores [10],[11],[12] that are not calibrated against cohort data; conditions are self-declared. (5) *Level resolution.* Very Unhealthy and Hazardous are merged into one interface level, so AQI above 300 is not distinguished in advisories. (6) *Data provenance.* Observations come from a third-party aggregator at a single point whose calibration we do not control, and only instantaneous readings are reported. (7) *Credential handling.* The data-provider credential is currently a single configured key; production deployment should use secret storage.

C. Future Work

Planned work includes: connecting the Open-Meteo eight-hour ozone average and a 24-hour SO2 average to the live pipeline; regenerating training data from the 2024 breakpoints with correlated source-profile mixtures (traffic, industrial, biomass) and validating against archived sensor data; using clustering for source attribution rather than severity; replacing the additive risk structure with a gated or multiplicative form that removes the clean-air floor; adding LSTM-based 24–72 hour forecasting [7]; conducting usability studies with at-risk groups [11]; and adding multilingual support [1].

X. CONCLUSION

This paper presented Aerocare, the second-generation AirSense system. The AQI engine now implements the six-pollutant US EPA calculation with the 2024 PM2.5 breakpoints, pollutant-specific truncation, and SO2/O3 averaging-period logic, and reproduces all 78 tabulated breakpoint endpoints. The 2024 revision changes the severity level of 5.6% of the 0–500 µg/m³ PM2.5 range and of 25% of the 100–200 µg/m³ range, which matters to any platform still using legacy breakpoints. The six-feature K-Means model shows moderate cluster validity (silhouette 0.48, Calinski–Harabasz 245.0, Davies–Bouldin 0.84) but does not outperform the six-pollutant formula as a severity classifier; its appropriate role is composition profiling. The condition-aware risk engine was enumerated in full (65 combinations), revealing a baseline floor for the most vulnerable conditions that future versions should remove. By reporting both the gains and the negative results, the paper aims to give practitioners a calibrated view of what regulatory fidelity, unsupervised learning, and personalized scoring each contribute to air quality intelligence.

ACKNOWLEDGMENT

The authors thank Mr. Ritesh Kumar for his 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] M. Mendez, M. G. Merayo, and M. Nunez, "Machine learning algorithms to forecast air quality: a survey," Artificial Intelligence Review, vol. 56, pp. 10031-10066, 2023, doi: 10.1007/s10462-023-10424-4.
[7] S. Zhou, W. Wang, L. Zhu, Q. Qiao, and Y. Kang, "Deep-learning architecture for PM2.5 concentration prediction: A review," Environmental Science & Ecotechnology, vol. 21, art. 100400, 2024, doi: 10.1016/j.ese.2024.100400.
[8] Z. Zhang, S. Zhang, C. Chen, and J. Yuan, "A systematic survey of air quality prediction based on deep learning," Alexandria Engineering Journal, vol. 93, pp. 128-141, 2024, doi: 10.1016/j.aej.2024.03.031.
[9] F. N. A. Suris, M. A. A. Bakar, N. M. Ariff, M. S. M. Nadzir, and K. Ibrahim, "Malaysia PM10 air quality time series clustering based on dynamic time warping," Atmosphere, vol. 13, no. 4, art. 503, 2022, doi: 10.3390/atmos13040503.
[10] H. Boogaard et al., "Long-term exposure to traffic-related air pollution and selected health outcomes: A systematic review and meta-analysis," Environment International, vol. 164, art. 107262, 2022, doi: 10.1016/j.envint.2022.107262.
[11] G. Markozannes et al., "Outdoor air quality and human health: An overview of reviews of observational studies," Environmental Pollution, vol. 306, art. 119309, 2022, doi: 10.1016/j.envpol.2022.119309.
[12] J. de Bont, S. Jaganathan, M. Dahlquist, A. Persson, M. Stafoggia, and P. Ljungman, "Ambient air pollution and cardiovascular diseases: An umbrella review of systematic reviews and meta-analyses," Journal of Internal Medicine, vol. 291, no. 6, pp. 779-800, 2022, doi: 10.1111/joim.13467.
[13] V. Kumar, M. Sahu, and P. Biswas, "Source apportionment of particulate matter by application of machine learning clustering algorithms," Aerosol and Air Quality Research, vol. 22, art. 210240, 2022, doi: 10.4209/aaqr.210240.
[14] M. Gagolewski, M. Bartoszuk, and A. Cena, "Are cluster validity measures (in)valid?," Information Sciences, vol. 581, pp. 620-636, 2021, doi: 10.1016/j.ins.2021.10.004.
[15] A. Treuille, T. Teixeira, and T. Hammouda, "Streamlit: The fastest way to build data apps," Streamlit Inc., 2019.
[16] L. Giri, S. Jain, N. Gupta, and Meenu, "AirSense: A Real-Time Monitoring Air Quality Intelligence System," International Journal for Multidisciplinary Research (IJFMR), vol. 8, no. 3, May-June 2026, paper ID IJFMR260377670.
[17] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
[18] A. M. Ikotun, F. Habyarimana, and A. E. Ezugwu, "Cluster validity indices for automatic clustering: A comprehensive review," Heliyon, vol. 11, no. 2, art. e41953, 2025, doi: 10.1016/j.heliyon.2025.e41953.
[19] R. Todeschini, D. Ballabio, V. Termopoli, and V. Consonni, "Extended multivariate comparison of 68 cluster validity indices: A review," Chemometrics and Intelligent Laboratory Systems, vol. 251, art. 105117, 2024, doi: 10.1016/j.chemolab.2024.105117.
[20] K. N. Ismail, A. Seman, and K. A. F. Abu Samah, "A comparison between external and internal cluster validity indices," in Proc. IEEE 11th International Conference on System Engineering and Technology (ICSET), 2021, pp. 229-233.
[21] Open-Meteo, "Air Quality API documentation," https://open-meteo.com/en/docs/air-quality-api.
[22] OpenWeather, "Air Pollution API and Geocoding API documentation," https://openweathermap.org/api.