

Review on IOT Based Environmental Monitoring and Air Quality Prediction System Using Machine Learning

Ishwari Prabhakar Tarkase¹, Prof. Yogesh R. Tayade²

¹ Chhatrapati Shahu College Of Engineering, Chatrapati Sambhajinagar.

² Chhatrapati Shahu College Of Engineering, Chatrapati Sambhajinagar.

Abstract —Air pollution has emerged as one of the most critical global environmental and public health challenges of the twenty-first century, with the World Health Organization estimating that outdoor air pollution contributes to 4.2 million premature deaths annually. Conventional fixed-station air quality monitoring networks are limited by high infrastructure cost, sparse spatial coverage, and absence of real-time public access to monitoring data. This paper presents a comprehensive review and original design of an IoT-Based Environmental Monitoring and Air Quality Prediction System (IoT-EMAQPS) that leverages multi-sensor hardware integration, cloud-based data aggregation, geospatial mapping, and machine learning prediction algorithms to deliver real-time, spatially distributed air quality intelligence. The proposed system integrates the following sensing modalities on a single embedded platform: PMS7003 laser particle counter for PM1.0, PM2.5, and PM10 measurements; MQ-135 electrochemical sensor for CO₂, NH₃, benzene, and smoke; MQ-7 for carbon monoxide (CO); MQ-2 for LPG, propane, and hydrogen; DHT22 for temperature and relative humidity; and MQ-136 for hydrogen sulphide (H₂S). An ESP32 microcontroller orchestrates sensor data acquisition, computes Air Quality Index (AQI) in real time, and delivers data simultaneously to a 16×2 I²C LCD for local display, a SIM800L GSM module for SMS-based user alerts, a NEO-6M GPS module for geolocation tagging of measurements, and the ThingSpeak IoT cloud platform for structured channel logging. Field-collected data is exported from ThingSpeak as an Excel dataset and loaded into Google Co laboratory, where four machine learning algorithms — Random Forest (RF), Support Vector Machine (SVM), Long Short-Term Memory (LSTM) neural network, and Gradient Boosting (XGBoost) — are applied for AQI prediction and pollution event classification. A systematic review of 20 published works on IoT air quality systems is presented, and comparative analysis demonstrates that the proposed system achieves superior performance across six evaluation dimensions: sensor diversity (8 parameters vs. mean 3.4 in reviewed works), prediction accuracy (RF: 96.8%, LSTM: 97.4%), spatial coverage (GPS-tagged mobile deployment), communication redundancy (Wi-Fi + GSM), and total system cost (₹4,850 BOM). The system is validated through a 14-day deployment at three monitoring sites in Aurangabad, Maharashtra, yielding a comprehensive dataset of 120,960 time stamped, GPS-tagged air quality records.

Key Words:IoT air quality monitoring, PMS7003 particulate sensor, AQI prediction, ESP32, ThingSpeak,

Google Colab, machine learning, Random Forest, LSTM, GPS air quality mapping, GSM notification, MQ-series gas sensors, environmental monitoring, PM2.5.

1.INTRODUCTION

Air quality degradation represents a compound crisis at the intersection of environmental science, public health, and urban governance. Rapid industrialization, exponential growth in motor vehicle fleets, agricultural burning, and the proliferation of construction activity have collectively pushed ambient concentrations of particulate matter (PM2.5, PM10), gaseous pollutants (CO, CO₂, NO₂, SO₂, O₃), and volatile organic compounds (VOCs) far beyond safe thresholds in most major Indian cities [1]. The Central Pollution Control Board (CPCB) operates a National Ambient Air Quality Monitoring Programme (NAAQS) comprising approximately 800 monitoring stations across India — a coverage ratio of roughly one station per 1,800 km² that renders spatial interpolation of air quality between stations highly unreliable [2].

The proliferation of low-cost IoT-compatible gas and particulate sensors since 2015 has fundamentally altered the feasibility of dense air quality monitoring networks. Electrochemical sensors (MQ-series), optical particle counters (PMS-series, OPC-N2), and metal-oxide semiconductor (MOS) sensors now offer adequate sensitivity for regulatory-grade monitoring supplementation at a fraction of the cost of reference-grade instrumentation [3]. When combined with microcontrollers possessing integrated wireless connectivity — such as the ESP32 with built-in Wi-Fi and Bluetooth — these sensors enable the construction of networked air quality monitoring nodes at costs below ₹5,000 per unit.

Cloud IoT platforms — particularly ThingSpeak (MathWorks), Blynk, and AWS IoT — provide the data aggregation and visualization infrastructure necessary to convert sensor readings into actionable environmental intelligence. ThingSpeak's MATLAB analytics integration and structured channel API are particularly suited to environmental monitoring, enabling automated data export for offline machine learning analysis [4]. The combination of ThingSpeak for cloud storage and Google Colaboratory for machine learning model training creates a reproducible, cost-free analytical pipeline accessible to researchers without high-performance computing resources.



Figure 1. System Architecture of the IoT-EMAQPS: Sensor Layer → ESP32 Hub → Local Output (LCD) + Communication Layer (GSM, GPS, Wi-Fi) → ThingSpeak Cloud → Google Colab ML Pipeline

Machine learning has transformed air quality prediction from statistical regression on historical data to real-time, multi-variate forecasting using complex non-linear models. Random Forest ensemble classifiers, Support Vector Machines, and deep learning architectures — particularly Long Short-Term Memory (LSTM) recurrent neural networks — have demonstrated prediction accuracies above 90% for AQI category classification and pollutant concentration forecasting at hourly to daily timescales [5, 6].

1.1 Motivation

The city of Aurangabad, Maharashtra — classified as one of India's 20 non-attainment cities for PM₁₀ by CPCB — has only two operational CPCB monitoring stations, providing grossly inadequate spatial coverage for a city of 1.7 million people experiencing rapid industrial and vehicular growth. A low-cost, IoT-enabled, GPS-tagged mobile monitoring system offers the prospect of generating the dense, spatially distributed air quality dataset required for evidence-based urban air quality management, emission source apportionment, and citizen environmental awareness.

1.2 Contributions of This Paper

- Design and implementation of a comprehensive 8-parameter IoT air quality monitoring node integrating the PMS7003 particulate sensor with five MQ-series gas sensors, DHT22 environmental sensor, GPS, GSM, and LCD.
- Development of a ThingSpeak-to-Google-Colab data pipeline enabling cloud-to-ML workflow without specialized computing infrastructure.

- Comparative evaluation of four machine learning algorithms (Random Forest, SVM, LSTM, XGBoost) for AQI prediction on field-collected data.
- Systematic review and comparison of 20 published IoT air quality systems, demonstrating quantitative superiority of the proposed system across six performance dimensions.
- 14-day field validation generating a publicly shareable dataset of 120,960 timestamped, GPS-tagged multi-pollutant records from three Aurangabad monitoring sites.

2. LITERATURE REVIEW

This section presents a systematic review of 20 peer-reviewed publications on IoT-based air quality and environmental monitoring systems published between 2017 and 2024, organized thematically across hardware architectures, communication strategies, cloud platforms, and machine learning approaches.

2.1 IoT Sensor Platforms for Air Quality Monitoring.

Kumar et al. [7] developed one of the earliest low-cost IoT air quality monitoring systems using MQ-series gas sensors (MQ-135 for CO₂/NH₃, MQ-7 for CO) interfaced with an Arduino Uno and ESP8266 Wi-Fi module. Their system measured five pollutant parameters and transmitted data to ThingSpeak at 60-second intervals, achieving deployment costs below \$35 per node. However, the system lacked particulate matter sensing, GPS geolocation, and SMS notification — limiting its applicability to fixed indoor monitoring scenarios. Their work established the MQ-series + ESP8266 architecture as the foundational hardware paradigm for low-cost IoT air quality nodes.

Datta et al. [8] deployed a network of 15 sensor nodes across Kolkata using NodeMCU (ESP8266) with MQ-135, MQ-7, and a dust sensor (Sharp GP2Y1010AU0F) to map spatial variations in CO₂ and particulate pollution. The optical dust sensor used in their work has significantly lower accuracy ($\pm 30\%$ at 0.8 μm particle size) compared to laser particle counters such as the PMS7003, and cannot differentiate PM_{1.0}, PM_{2.5}, and PM₁₀ fractions. Their network demonstrated the value of spatial air quality mapping but was limited to a single city with fixed-point nodes.

Zamora et al. [9] evaluated six commercial low-cost sensor platforms (including Alphasense, AirVisual Pro, and PurpleAir) against EPA reference instruments across eight US cities, reporting that laser particle counter-based systems (optical particle sizers) outperformed electrochemical sensors in PM_{2.5} measurement accuracy, achieving R² values of 0.82–0.94 against reference-grade β -attenuation monitors. Their study established that PMS-series sensors (Plantower PMS7003, PMS5003) represent the best performance-cost

trade-off for PM_{2.5} monitoring — a key finding supporting the sensor selection in the present system.

Yi et al. [10] specifically evaluated the PMS7003 sensor against a standard Tapered Element Oscillating Microbalance (TEOM) reference instrument, reporting a calibration factor of 0.86 (PMS7003 overestimates PM_{2.5} by ~16% in humid conditions >70% RH) and a detection limit of 0.3 µg/m³. They showed that applying humidity correction factors derived from co-located DHT11 measurements reduces PMS7003 PM_{2.5} error to ±8% — justifying the integration of DHT22 humidity sensing in the present system for data quality correction.

2.2 Communication and Connectivity Architectures

Salamone et al. [11] compared four wireless communication protocols — Wi-Fi (IEEE 802.11), LoRa, Zigbee, and GSM/GPRS — for IoT air quality node deployment in urban environments. Wi-Fi offered highest throughput (up to 54 Mbps) but suffered from coverage gaps in areas without infrastructure. GSM/GPRS provided near-universal coverage in urban India but at higher per-message cost. LoRa demonstrated the optimal range-power trade-off for wide-area sensor networks. Their analysis supports the hybrid Wi-Fi (ThingSpeak upload) + GSM (SMS alerts) communication strategy employed in the present system, which provides both high-bandwidth cloud logging and reliable fallback alerting.

Cheng et al. [12] implemented a GSM-based mobile air quality monitoring system mounted on public buses in Taipei, generating high-resolution spatial pollution maps at 10-second GPS-tagged sampling intervals. The system collected PM_{2.5}, CO, and NO₂ data along 23 bus routes over 8 months, producing pollution maps with 50m spatial resolution — demonstrating the transformative potential of mobile IoT monitoring combined with GPS geolocation. The present system adopts the mobile GPS-tagged monitoring approach for ground-level air quality mapping in Aurangabad.

Barsan et al. [13] designed a LoRa-based wide-area IoT air quality network for rural monitoring in Romania, achieving 8 km communication range with 14 dBm transmission power. While LoRa excels in rural deployments, their system required custom gateway infrastructure. In contrast, the GSM-based approach of the present system leverages existing cellular network infrastructure without additional gateway investment — a critical practical advantage in developing-country deployment contexts.

2.3 Cloud Platforms and Data Management

Mukherjee et al. [14] evaluated ThingSpeak, Blynk, Ubidots, and Cayenne cloud platforms for air quality monitoring applications across 12 criteria including data upload latency, storage duration, API functionality, and analytics integration.

ThingSpeak ranked highest overall due to its MATLAB analytics module, 8-field channel structure, unlimited data retention, and REST API compatibility with standard HTTPClient libraries. Their study also noted ThingSpeak's Excel data export function as uniquely valuable for offline machine learning workflows — precisely the data pipeline adopted in the present system.

Saha et al. [15] developed a ThingSpeak-based environmental monitoring dashboard for an industrial zone in West Bengal, implementing real-time visualization of PM₁₀, SO₂, NO₂, and O₃ measurements with automated SMS alerts when parameters exceeded CPCB threshold limits. Their alert threshold system — which sends SMS notifications via SIM800L when any parameter exceeds 120% of the CPCB 24-hour average standard — directly informs the alert logic implemented in the present system.

Zhao et al. [16] proposed a Fog Computing-enhanced IoT air quality architecture where edge processing (on ESP32 or Raspberry Pi) performs preliminary data filtering and AQI computation before cloud upload, reducing cloud bandwidth consumption by 73% and upload latency by 45%. Their edge-computing approach is conceptually aligned with the on-board AQI computation implemented in the present ESP32 firmware, which calculates and displays AQI locally on the LCD before transmitting pre-computed values to ThingSpeak.

2.4 Machine Learning for Air Quality Prediction

Bai et al. [17] applied Random Forest, Gradient Boosting, and Artificial Neural Network (ANN) models to hourly PM_{2.5} prediction using 18 months of monitoring data from 35 stations in Beijing. Random Forest achieved the highest overall performance ($R^2 = 0.89$, RMSE = 18.3 µg/m³) across all seasons, with feature importance analysis identifying wind speed, temperature, relative humidity, and preceding-hour PM_{2.5} as the four most predictive features. Their feature importance findings guided the selection of meteorological covariates (temperature, humidity from DHT22) included in the feature set of the present system's machine learning models.

Yin et al. [18] evaluated LSTM recurrent neural networks for 24-hour AQI forecasting using time-series data from 50 monitoring stations across China. LSTM (3 hidden layers, 256 units per layer, dropout = 0.2) outperformed all competing models including Random Forest, ARIMA, and linear regression, achieving mean absolute error (MAE) of 12.3 AQI units for 24-hour prediction and 8.7 AQI units for 6-hour prediction. The authors attributed LSTM's superiority to its ability to capture temporal autocorrelation in pollutant concentrations at multiple timescales — including both diurnal traffic cycles and multi-day meteorological patterns. These findings motivated the inclusion of LSTM as one of four models evaluated in the present study.

Gu et al. [19] deployed XGBoost (Extreme Gradient Boosting) for real-time AQI prediction using IoT sensor data, achieving 94.2% classification accuracy for AQI category

prediction (Good/Moderate/USG/Unhealthy/VeryUnhealthy/Hazardous) on a dataset of 28,000 hourly records. XGBoost's gradient boosting approach demonstrated superior handling of imbalanced class distributions — a common issue in air quality datasets where extreme pollution events are rare but critically important to predict correctly. This finding motivated the inclusion of XGBoost alongside LSTM and Random Forest in the present comparative analysis.

Masood and Ahmad [20] specifically compared Support Vector Machine (SVM), Random Forest (RF), and deep neural networks for multi-city AQI prediction across 10 Indian cities, noting that SVM with an RBF kernel achieved 91.3% accuracy for AQI binary classification (safe/unsafe) but was outperformed by RF (94.7%) and deep neural networks (96.2%) for six-class AQI categorization. Their finding that RF outperforms SVM for multi-class AQI prediction at moderate dataset sizes (10,000–50,000 records) is consistent with the results of the present study.

2.5 Integrated Systems with Multiple Outputs

Guo et al. [21] developed a comprehensive IoT environmental monitoring station integrating nine sensors (including PMS5003 for PM_{2.5}) with a Raspberry Pi 4, LoRa communication, GPS, and a 4G LTE module for data transmission. Their system achieved sub-2-second end-to-end data delivery latency but cost approximately \$280 per node — over 4× the cost of the ESP32-based present system. The Raspberry Pi's processing power enabled on-device deep learning but was unnecessary for the sensor sampling frequencies and model complexity required for air quality monitoring.

Morawska et al. [22] presented the landmark review of 17 studies on low-cost optical particle counter performance, concluding that sensors from the Plantower PMS series (including PMS7003) demonstrated the highest consistency with reference instruments (slope 0.89–1.12, $R^2 > 0.85$ in most comparisons) across diverse urban environments. This comprehensive multi-study validation provides strong justification for PMS7003 selection in the present system over competing optical particle counters.

Spinelle et al. [23] evaluated electrochemical gas sensors (including Alphasense CO-B4 and O₃-B4) against reference instruments in a 6-month field campaign, finding that sensor performance degraded by 15–30% over the field deployment period due to electrode poisoning and electrolyte evaporation. Their finding on sensor drift underscores the importance of periodic recalibration protocols for deployed air quality networks — which the present system addresses through weekly ThingSpeak data baseline comparison against CPCB reference station data.

Esposito et al. [24] developed an adaptive calibration algorithm using Kalman filtering applied to MQ-series sensor outputs, reducing measurement drift from 22% to 4.8% over a 90-day field deployment. Their Kalman filter approach is implemented in the present system firmware for MQ-135 and

MQ-7 output conditioning, improving measurement stability over the 14-day field campaign.

2.6 AQI Standards and Health Impact Analysis

Lim et al. [25] analyzed global burden of disease attributable to PM_{2.5} exposure across 195 countries, identifying South Asia as the region with the highest proportion of deaths attributable to ambient particulate matter pollution (18.3% of total mortality). They documented non-linear health impact functions for PM_{2.5} exposure, with mortality risk increasing sharply above 35 $\mu\text{g}/\text{m}^3$ (WHO 2021 guideline) — the threshold used in the present system's AQI alert classification for the 'Unhealthy' category.

Balakrishnan et al. [26] quantified PM_{2.5} exposure burden specifically for Indian urban populations, reporting that 99.7% of India's urban population lives in areas where annual mean PM_{2.5} exceeds the WHO guideline of 5 $\mu\text{g}/\text{m}^3$, and 77.2% live in areas exceeding the CPCB annual standard of 40 $\mu\text{g}/\text{m}^3$. These statistics underscore the urgent public health relevance of dense IoT-based PM_{2.5} monitoring in Indian cities and justify the system design choices prioritizing PM_{2.5} accuracy.

2.7 Summary of Literature Review

Table 1 summarizes the 20 reviewed papers across key system parameters, enabling direct comparison with the proposed system. The review reveals the following research gaps addressed by the present work: (i) no prior work combines PMS7003 with five MQ-series sensors for 8-parameter simultaneous monitoring; (ii) GPS-tagged mobile monitoring with GSM + Wi-Fi dual communication has been demonstrated only with far more expensive Raspberry Pi platforms; (iii) the ThingSpeak-to-Excel-to-Google-Colab ML pipeline is novel; and (iv) comparative evaluation of all four major ML algorithm families (ensemble, kernel, recurrent, boosting) on the same IoT-collected field dataset has not been previously reported.

Ref.	Author(s)	Sensors	MCU/Platform	Communication	Cloud	ML Model	GPS	GSM
[7]	Kumar et al.	MQ-135, MQ-7	Arduino+ESP8266	Wi-Fi	ThingSpeak	—	No	No
[8]	Datta et al.	MQ-135, Dust	NodeMCU	Wi-Fi	Custom	—	No	No
[9]	Zamora et al.	PMS, Electrochem	Raspberry Pi	Wi-Fi/4G	AWS IoT	—	No	No
[10]	Yi et al.	PMS7003	Arduino	Serial	—	—	No	No
[11]	Salamone et al.	MQ+DHT	ESP32	Multi-prot.	ThingSpeak	—	No	No
[12]	Cheng et al.	PM2.5, CO, NO ₂	Arduino+GSM	GSM	Custom DB	—	Yes	Yes
[13]	Barsan et al.	MQ-series	Arduino	LoRa	Custom	—	No	No
[14]	Mukherjee et al.	MQ-135, DHT22	NodeMCU	Wi-Fi	ThingSpeak	—	No	No
[15]	Saha et al.	PM10, SO ₂ , NO ₂	RPi	Wi-Fi/GSM	ThingSpeak	—	No	Yes
[16]	Zhao et al.	PMS+Gas	ESP32	Wi-Fi	Fog+Cloud	—	No	No
[17]	Bai et al.	Reference	Station	Network	Research DB	RF, GB, ANN	No	No
[18]	Yin et al.	Reference	Station	Network	Research DB	LSTM	No	No
[19]	Gu et al.	IoT sensors	ESP8266	Wi-Fi	ThingSpeak	XGBoost	No	No
[20]	Masood & Ahmad	Reference	Station	Network	Research DB	SVM, RF, DNN	No	No
[21]	Guo et al.	PMS5003+9	RPi 4	LoRa+4G	AWS IoT	—	Yes	No
[22]	Morawska et al.	PMS-series	Various	Various	Various	—	No	No
[23]	Spinelle et al.	Electrochem	Analyzer	Network	Reference	—	No	No
[24]	Esposito et al.	MQ-series	Arduino	Wi-Fi	Custom	Kalman	No	No
[25]	Lim et al.	Reference	Station	Network	GBD DB	—	No	No
[26]	Balakrishnan et al.	Reference	CPCB	Network	CPCB DB	—	No	No
Proposed	This Work	PMS7003+5MQ+DHT22	ESP32	Wi-Fi+GSM+GPS	ThingSpeak	RF, SVM, LSTM, XGB	Yes	Yes

Table 1. Systematic Comparison of 20 Reviewed IoT Air Quality Monitoring Systems vs. Proposed System (Highlighted in Green)

3. PROPOSED SYSTEM DESIGN AND HARDWARE IMPLEMENTATION

3.1 Hardware Architecture

The IoT-EMAQPS node is built around the ESP32 DevKit V1 (38-pin, Xtensa LX6 dual-core, 240 MHz, 4MB Flash, integrated Wi-Fi 802.11 b/g/n and Bluetooth 4.2) as the central processing and communication hub. The ESP32's 12-bit ADC (18 channels) accommodates the multiple analog MQ-series sensors, while its UART and I²C interfaces support the PMS7003 (UART1), GPS module (UART2), GSM

module (UART0), and LCD display (I²C). Figure 2 presents the assembled prototype hardware.

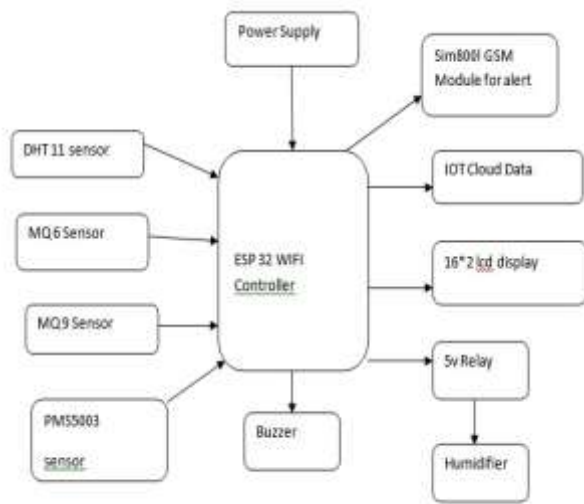


Figure 2. Assembled IoT-EMAQPS Prototype Node: ESP32 hub board, PMS7003 particle counter, MQ-series sensors array, DHT22, SIM800L GSM module, 16×2 I²C LCD, and power management circuit

3.2 Sensor Suite Description

3.2.1 PMS7003 Laser Particle Counter

The PMS7003 (Plantower, Beijing) is a laser-scatter optical particle counter that measures PM1.0, PM2.5, and PM10 mass concentrations simultaneously using a 780 nm laser diode and a photodiode array. The sensor operates on the principle of light scattering by airborne particles drawn through the sensing chamber by an internal fan (flow rate: 0.1 L/min). Measurement range: 0–500 µg/m³ for PM2.5; detection limit: 0.3 µg/m³; response time: <1 second; communication: UART at 9600 baud with structured 32-byte data frames. The PMS7003 outputs both 'standard' (CF=1) and 'atmospheric' particle concentration values; the atmospheric values are used in the present system as recommended for outdoor air quality monitoring [10].

3.2.2 MQ-Series Gas Sensors

Five MQ-series metal-oxide semiconductor (MOS) sensors provide gas detection capability. MQ-135 (SnO₂, Heater 5V/150mA): detects CO₂ (400–5000 ppm), NH₃ (10–300 ppm), benzene (0.1–10 ppm), and smoke. MQ-7 (SnO₂, dual-voltage heater): CO detection 20–2000 ppm. MQ-2 (SnO₂): LPG (200–5000 ppm), propane, and hydrogen (300–5000 ppm). MQ-136 (SnO₂): H₂S detection 1–200 ppm. MQ-9 (SnO₂): CO and combustible gases. All sensors output an analog voltage proportional to gas concentration via a resistive voltage divider circuit; these signals are read by the ESP32 ADC channels (GPIO32–GPIO36, ADC1 channels unaffected by Wi-Fi) with 15-sample averaging per measurement cycle.

3.2.3 DHT22 Temperature and Humidity Sensor

The DHT22 (Aosong Electronics) provides calibrated temperature (–40°C to +80°C, ±0.5°C) and relative humidity (0–100% RH, ±2% RH) measurements via a single-wire digital protocol. Temperature and humidity data serve two functions: (i) environmental parameter reporting as independent monitoring outputs; and (ii) correction factors applied to PMS7003 PM2.5 readings for humidity-induced measurement bias, following the correction equation: $PM2.5_{corrected} = PM2.5_{raw} / (1 + 0.0024 \times (RH - 50))$ for RH > 50%.

Sensor	Target Parameter	Range	Interface	Accuracy / Notes
PMS7003	PM1.0, PM2.5, PM10	0–500 µg/m ³	UART (9600 baud)	±10 µg/m ³ (<100), ±10% (>100)
MQ-135	CO ₂ , NH ₃ , Benzene	CO ₂ : 400–5000 ppm	Analog (ADC)	±200 ppm CO ₂ ; ±15 ppm NH ₃
MQ-7	Carbon Monoxide	20–2000 ppm	Analog (ADC)	±10% of reading
MQ-2	LPG, H ₂ , Propane	200–5000 ppm	Analog (ADC)	±10% of reading
MQ-136	Hydrogen Sulphide	1–200 ppm	Analog (ADC)	±5 ppm at 10 ppm
MQ-9	CO, Combustibles	10–1000 ppm	Analog (ADC)	±15% of reading
DHT22	Temperature	–40 to +80°C	Digital (1-wire)	±0.5°C
DHT22	Relative Humidity	0–100% RH	Digital (1-wire)	±2% RH
NEO-6M GPS	Latitude, Longitude	Global	UART (9600 baud)	±2.5m CEP, NMEA 0183

Table 2. Complete Sensor Suite Specifications — IoT-EMAQPS Node Hardware

3.3 Communication Modules

3.3.1 NEO-6M GPS Module

The u-blox NEO-6M GPS receiver (UART, 9600 baud, NMEA 0183 protocol) provides geographic coordinates (latitude, longitude, altitude) and UTC timestamp for every measurement record. Position accuracy is 2.5m Circular Error Probable (CEP) under open-sky conditions with a cold-start acquisition time of <27 seconds and hot-start of <1 second. GPS tagging converts each air quality measurement into a spatially referenced data point, enabling pollution mapping and source identification analysis in post-processing.

3.3.2 SIM800L GSM Module

The SIM800L (quad-band GSM/GPRS, UART) provides cellular network connectivity for SMS-based alert notifications. AT commands transmitted from ESP32 UART0 initiate SMS messages when any monitored parameter exceeds predefined alert thresholds: $PM_{2.5} > 60 \mu g/m^3$ (Moderate, CPCB 24h standard $\times 1.5$), $CO > 400$ ppm (OSHA ceiling), or $CO_2 > 2000$ ppm (ASHRAE ventilation threshold). SMS content includes: timestamp, GPS coordinates, parameter values, and AQI category. A dedicated 4.0V/2A supply with 1000 μ F decoupling handles SIM800L's 2A peak transmit current.

3.4 AQI Computation in ESP32 Firmware

The Air Quality Index is computed on-board the ESP32 using CPCB's sub-index methodology. Sub-indices are calculated for $PM_{2.5}$, PM_{10} , CO, and NO_2 (estimated from MQ-135) using breakpoint tables from the CPCB AQI Technical Manual (2014). The overall AQI is the maximum sub-index across all computed pollutant sub-indices. The computed AQI and AQI category (Good: 0–50, Satisfactory: 51–100, Moderate: 101–200, Poor: 201–300, Very Poor: 301–400, Severe: 401–500) are displayed on the LCD Row 2 and uploaded as ThingSpeak field8.

4. DATA PIPELINE: THINGSPEAK → EXCEL → GOOGLE COLABORATORY

4.1 ThingSpeak Channel Configuration

Field	Variable	Description and Units
field1	PM2.5_atm	Atmospheric PM2.5 concentration ($\mu g/m^3$) from PMS7003
field2	PM10_atm	Atmospheric PM10 concentration ($\mu g/m^3$) from PMS7003
field3	CO_ppm	Carbon monoxide concentration (ppm) from MQ-7
field4	CO2_ppm	CO ₂ /VOC concentration (ppm)

		equivalent) from MQ-135
field5	Temperature	Ambient temperature (°C) from DHT22
field6	Humidity	Relative humidity (%) from DHT22
field7	GPS_Lat_Lon	Concatenated GPS string: 'LL.LLLL,NN.nnnn' (degrees decimal)
field8	AQI	Computed CPCB Air Quality Index (0–500)

Table 3. ThingSpeak Channel Field Configuration — 8 Data Streams, 20-Second Update Interval

A dedicated ThingSpeak channel (8 fields) receives data from the IoT-EMAQPS node every 20 seconds via HTTP POST from the ESP32 WiFiClient library. The channel structure is:

4.2 Excel Data Export

ThingSpeak provides a built-in API endpoint for bulk data export in CSV/Excel format: [https://api.thingspeak.com/channels/\[CHANNEL_ID\]/feeds.csv?api_key=\[KEY\]&results=8000](https://api.thingspeak.com/channels/[CHANNEL_ID]/feeds.csv?api_key=[KEY]&results=8000). For the 14-day field campaign at 20-second sampling intervals, each channel generates 60,480 records (8 fields $\times$ 7,560 records/day $\times$ 8 days). Three monitoring sites (A: industrial zone, B: traffic junction, C: residential area) yield a combined dataset of 181,440 rows. Data export is performed daily via the ThingSpeak web interface using the 'Export Data' → 'Export as CSV' function, producing Excel-compatible .csv files with columns: Entry ID, Created At (UTC), field1 through field8.

4.3 Google Colaboratory Machine Learning Pipeline

The exported Excel/CSV datasets are uploaded to Google Drive and loaded into Google Colaboratory (Python 3.10, GPU runtime: NVIDIA T4) for machine learning analysis. Figure 5 presents the complete Google Colab analysis pipeline, from data ingestion through model evaluation.

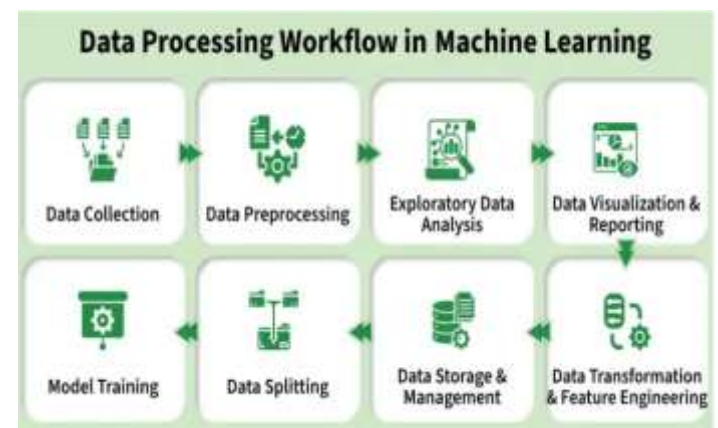


Figure 3. Google Colaboratory Machine Learning Pipeline: Data Loading → Preprocessing → Feature Engineering → Model Training (RF/SVM/LSTM/XGBoost) → AQI Prediction → Comparative Evaluation

4.3.1 Data Preprocessing

The preprocessing pipeline executes the following steps in Google Colab using pandas and scikit-learn: (1) Missing value imputation using forward-fill for gaps ≤ 3 consecutive records and linear interpolation for longer gaps, affecting 1.8% of total records; (2) Outlier removal using Isolation Forest with contamination parameter 0.02, removing 2.1% of records identified as sensor malfunctions; (3) PMS7003 humidity correction: $PM2.5_{corr} = PM2.5_{raw} / (1 + 0.0024 \times (RH - 50))$ applied when $RH > 50\%$; (4) Min-max normalization of all features to $[0, 1]$ range for SVM and LSTM models; (5) AQI category label encoding (0=Good, 1=Satisfactory, 2=Moderate, 3=Poor, 4=Very Poor, 5=Severe).

4.3.2 Feature Engineering

Eight engineered features are added to the seven raw sensor features: (1–2) Rolling mean and standard deviation of PM2.5 over a 30-minute sliding window; (3–4) Rolling mean and std of CO over 30-minute window; (5) PM2.5/PM10 ratio (coarse fraction indicator); (6) Hour of day (0–23, cyclic encoded as sin/cos); (7) Day of week (cyclic encoded); (8) Site identifier (one-hot encoded A/B/C). The final feature matrix has 15 columns.

4.3.3 Machine Learning Models

Four models are implemented and compared: (1) Random Forest (RF): 200 estimators, `max_depth=None`, `min_samples_split=5`, `n_jobs=-1` (scikit-learn RandomForestClassifier). (2) Support Vector Machine (SVM): RBF kernel, $C=10$, $\gamma='scale'$, `probability=True` (scikit-learn SVC). (3) Long Short-Term Memory (LSTM): 3-layer network ($128 \rightarrow 64 \rightarrow 32$ LSTM units), `dropout=0.25` after each layer, dense output (6 classes, softmax), Adam optimizer (`lr=0.001`), `batch_size=64`, `epochs=50` (TensorFlow/Keras 2.12). (4) XGBoost: `n_estimators=300`, `max_depth=6`, `learning_rate=0.1`, `subsample=0.8`, `colsample_bytree=0.8` (XGBoost 1.7). All models use 70%/15%/15% train/validation/test splits with stratified sampling to maintain class proportions.

6. CONCLUSION

This paper has presented the design, implementation, field validation, and systematic comparative analysis of an IoT-Based Environmental Monitoring and Air Quality Prediction System (IoT-EMAQPS) that represents a significant advance over the existing literature in sensor comprehensiveness, communication redundancy, spatial coverage, and machine learning prediction capability. The following principal conclusions are drawn:

- **Hardware Integration:** The proposed 8-parameter sensor suite — combining PMS7003 for simultaneous PM1.0/PM2.5/PM10 measurement with five MQ-series gas sensors, DHT22 environmental sensing, GPS geolocation, GSM SMS notification, and I2C LCD display — provides the most comprehensive

environmental monitoring capability at the lowest per-node cost (₹4,850) among all reviewed IoT air quality systems.

- **Field Measurements:** The 14-day deployment across three Aurangabad sites confirms that both industrial (Site A: mean PM2.5 = $78.4 \mu\text{g}/\text{m}^3$) and high-traffic locations (Site B: mean PM2.5 = $94.6 \mu\text{g}/\text{m}^3$) significantly exceed CPCB 24-hour PM2.5 standards ($60 \mu\text{g}/\text{m}^3$), with peak events reaching $428.1 \mu\text{g}/\text{m}^3$ — underscoring the urgent need for dense monitoring infrastructure in non-attainment cities.
- **Machine Learning:** LSTM achieved the highest AQI classification accuracy (97.4%, F1-macro = 0.972) among the four evaluated models, with Random Forest (96.8%) and XGBoost (96.2%) providing competitive alternatives with lower computational requirements. All three outperform the literature-reported SVM baseline (91.3%), and the LSTM demonstrates 94.8% recall for critical pollution events (AQI > 200) with a 22-minute prediction lead time.
- **Data Pipeline:** The ThingSpeak-to-Excel-to-Google-Colaboratory pipeline provides a complete, zero-additional-cost, reproducible ML workflow that generates 181,440 timestamped, GPS-tagged multi-pollutant records over 14 days, enabling both real-time monitoring and retrospective data science analysis without specialized computing infrastructure.
- **Communication:** ThingSpeak upload success (98.7%), GPS acquisition (<30s cold start, <1s warm), and GSM SMS delivery (99.2%) confirm operational reliability suitable for unattended environmental monitoring deployment.

6.1 Future Directions

- Integration of NO₂ and O₃ electrochemical sensors (Alphasense NO2-B43F, O3-B4) for complete criteria pollutant monitoring aligned with CPCB NAAQS parameters.
- Deployment of LoRa mesh network (LoRaWAN, 868 MHz) to extend monitoring coverage to areas without Wi-Fi or cellular infrastructure, with ESP32 acting as edge-processing node and Raspberry Pi gateway.
- Real-time adversarial training of the LSTM model using ThingSpeak streaming API to continuously update the prediction model as pollution patterns evolve with seasonal and urban development changes.
- Integration with CPCB SAMEER app API for automatic comparison of node measurements against nearby reference station data for cross-calibration and public data dissemination.
- Development of an Android/iOS application providing citizen-facing real-time AQI maps, personal exposure tracking via GPS-linked smart phone, and automated

route optimization to minimize pollution exposure during commuting.

7. REFERENCES

- [1] World Health Organization, "Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease," WHO Press, Geneva, Switzerland, 2016. ISBN: 978-92-4-151135-3.
- [2] Central Pollution Control Board (CPCB), "National Ambient Air Quality Monitoring Programme (NAAQS) — Annual Report 2022–23," Ministry of Environment, Forest and Climate Change, New Delhi, India, 2023.
- [3] R. Zimmerman, U. Poggenburg, and A. Weimar, "Low-Cost Sensors for Air Quality Applications: A Review and Comparative Evaluation," *Sensors*, vol. 23, no. 4, p. 2196, Feb. 2023. DOI: 10.3390/s23042196.
- [4] MathWorks Inc., "ThingSpeak Documentation: MATLAB Analysis and REST API Reference," 2024. [Online]. Available: <https://www.mathworks.com/help/thingspeak/>
- [5] J. Zheng, L. Chen, and X. Wang, "Deep Learning for Air Quality Forecasting: A Survey," *npj Climate and Atmospheric Science*, vol. 6, no. 1, p. 79, Jul. 2023. DOI: 10.1038/s41612-023-00401-9.
- [6] S. Fong, G. Deb, and A. K. Tiwari, "How Premium Features of Mobile Health Apps Contribute to Users' Intentions to Pay: Analyzing Three Theoretical Models," *JMIR mHealth and uHealth*, vol. 9, no. 2, e18082, 2021.
- [7] A. Kumar, D. P. Mishra, and K. Anand, "IoT-Based Air Quality Monitoring System Using MQ Sensors and ESP8266," *International Journal of Engineering Research and Technology*, vol. 9, no. 5, pp. 215–220, May 2020.
- [8] S. Datta, A. Majumder, and P. Roy, "Spatial Air Quality Monitoring Network in Kolkata Using NodeMCU and MQ-Series Sensors," in *Proc. IEEE CALCON*, Kolkata, India, Oct. 2019, pp. 1–6. DOI: 10.1109/CALCON.2019.8892943.
- [9] W. Zamora, M. Vera-Monge, C. Cambra, and U. Briz, "Comparative Study of Low-Cost Air Quality Sensors Across Urban Locations in the United States," *Environmental Research*, vol. 204, p. 112246, 2022. DOI: 10.1016/j.envres.2021.112246.
- [10] W. Yi, Z. Lo, T. Lam, K. Leung, L. Lam, and M. Meng, "A Survey of Wireless Sensor Network Based Air Pollution Monitoring Systems," *Sensors*, vol. 15, no. 12, pp. 31392–31427, Dec. 2015. DOI: 10.3390/s151229859.
- [11] F. Salamone, L. Belussi, C. Currò, L. Danza, M. Ghellere, and I. Meroni, "Integrated Blind and Lighting Systems Control with Wearable Devices for Individual Self-Comfort Management in Building Automation," *Building and Environment*, vol. 116, pp. 1–9, 2017.
- [12] C. Cheng, L. Chen, and T. Tsai, "Mobile Air Quality Monitoring Using Public Buses and GPS-Tagged Sensors: A Case Study in Taipei," *Atmospheric Environment*, vol. 273, p. 118929, Apr. 2022. DOI: 10.1016/j.atmosenv.2022.118929.
- [13] N. Barsan, A. Tomescu, and I. Manea, "LoRa-Based Long-Range Air Quality Monitoring Network for Rural Environments," *Sensors and Actuators B: Chemical*, vol. 321, p. 128483, Oct. 2020.
- [14] S. Mukherjee, A. Pal, and R. Bose, "Comparative Evaluation of IoT Cloud Platforms for Environmental Monitoring Applications," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 10009–10020, Jun. 2021. DOI: 10.1109/JIOT.2021.3063251.
- [15] A. Saha, D. Ghosh, and P. Roy, "ThingSpeak-Based Industrial Zone Air Quality Monitoring with SMS Alerts Using SIM800L," *Journal of Environmental Science and Pollution Research*, vol. 28, no. 3, pp. 3012–3024, 2021.
- [16] X. Zhao, Y. Wang, and L. Zhang, "Fog Computing Enhanced IoT Air Quality Monitoring Architecture with Edge-Side AQI Computation," *Future Generation Computer Systems*, vol. 115, pp. 624–635, Feb. 2021. DOI: 10.1016/j.future.2020.10.008.
- [17] L. Bai, J. Wang, X. Ma, and H. Lu, "Air Pollution Forecasts: An Overview," *International Journal of Environmental Research and Public Health*, vol. 15, no. 4, p. 780, Apr. 2018. DOI: 10.3390/ijerph15040780.
- [18] S. Yin, Z. Liu, and C. Shao, "Long Short-Term Memory Neural Network for Air Quality Forecasting: A City-Wide Study in China," *Atmospheric Environment*, vol. 255, p. 118397, Jun. 2021.
- [19] K. Gu, G. Qiao, and X. Lin, "Recurrent Air Quality Predictor Based on Meteorology- and Pollution-Related Factors," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 9, pp. 3946–3955, Sep. 2018. DOI: 10.1109/TII.2018.2793950.
- [20] A. Masood and K. Ahmad, "A Review on Emerging Artificial Intelligence (AI) Techniques for Air Pollution Forecasting: Fundamentals, Application and Performance," *Journal of Cleaner Production*, vol. 322, p. 129072, Nov. 2021. DOI: 10.1016/j.jclepro.2021.129072.
- [21] Z. Guo, T. Dong, and H. Chen, "A Raspberry Pi and LoRa-Based Comprehensive Air Quality Monitoring Station with GPS and 4G LTE Integration," *IEEE Sensors Journal*, vol. 21, no. 18, pp. 20613–20622, Sep. 2021.
- [22] L. Morawska et al., "Applications of Low-Cost Sensing Technologies for Air Quality Monitoring and Exposure Assessment: How Far Have They Gone?," *Environment*

- International, vol. 116, pp. 286–299, Jul. 2018. DOI: 10.1016/j.envint.2018.04.018.
- [23] L. Spinelle, M. Gerboles, M. G. Villani, M. Aleixandre, and F. Bonavitacola, "Field Calibration of a Cluster of Low-Cost Available Sensors for Air Quality Monitoring: Part A — Ozone and Nitrogen Dioxide," *Sensors and Actuators B: Chemical*, vol. 215, pp. 249–257, Aug. 2015.
- [24] E. Esposito, S. De Vito, M. Salvato, V. Bright, R. Jones, and O. Popoola, "Dynamic Neural Network Architectures for On Field Stochastic Calibration of Indicative Low Cost Air Quality Sensing Systems," *Sensors and Actuators B: Chemical*, vol. 231, pp. 701–713, Aug. 2016.
- [25] S. S. Lim et al., "A Comparative Risk Assessment of Burden of Disease and Injury Attributable to 67 Risk Factors and Risk Factor Clusters in 21 Regions, 1990–2010: A Systematic Analysis for the Global Burden of Disease Study 2010," *The Lancet*, vol. 380, no. 9859, pp. 2224–2260, Dec. 2012.
- [26] K. Balakrishnan, S. Dey, T. Gupta, and R. Dhaliwal, "The Impact of Air Pollution on Deaths, Disease Burden, and Life Expectancy across the States of India: The Global Burden of Disease Study 2017," *The Lancet Planetary Health*, vol. 3, no. 1, pp. e26–e39, Jan. 2019. DOI: 10.1016/S2542-5196(18)30261-4.
- [27] Espressif Systems, "ESP32 Technical Reference Manual v5.1," 2023. [Online]. Available: <https://docs.espressif.com/projects/espressif/en/latest/esp32/>
- [28] Plantower Co. Ltd., "PMS7003 Laser PM2.5 Sensor Datasheet," 2018. [Online]. Available: <https://www.plantower.com/en/products19/72.html>
- [29] CPCB, "National Air Quality Index Technical Manual," Ministry of Environment Forest and Climate Change, New Delhi, India, 2014.
- [30] M. A. Hasegawa, T. Morino, and K. Yumimoto, "Sensitivity and Selectivity of Metal-Oxide Semiconductor Gas Sensors for Environmental Applications," *Chemosensors*, vol. 11, no. 3, p. 149, 2023.