Vol 8, Issue 5, 2025 **Impact Factor: 8.535** DOI: https://doi.org/10.31426/ijamsr.2025.8.5.8329

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

Real-Time Air Pollution Forecasting with IoT, KNN and Entropy in Crowded Areas

Sarita Jiyal

Research Scholar, Department of Computer Application,
Sri Satya Sai University of Technology & Medical Sciences, Sehore, M.P., India.

Dr. Rajendra Singh Kushwah

Research Supervisor, Department of Computer Application,
Sri Satya Sai University of Technology & Medical Sciences, Sehore, M.P., India.

ABSTRACT

Air pollution has emerged as a significant environmental and health issue, prompting the need for accurate prediction models to assess pollution levels in real time. This research proposes a novel framework named the Air Pollution Estimation Model (APEM), which utilizes the Internet of Things (IoT) to predict air pollution in crowded areas. The objectives include designing a low-cost, real-time monitoring system using gas sensors (MQ7 for CO and MQ135 for CO2) connected to an Arduino microcontroller, with data transmitted via WiFi and RF 433 modules. The methods involve data collection, K-Nearest Neighbors clustering, regression analysis, and entropy estimation based on Shannon's information gain theory to forecast pollution levels. Results from a dataset of 36,388 records collected from January to April 2020 demonstrate high accuracy, with mean absolute error (MAE) values ranging from 0.021 to 0.115 for various pollutants in 5-minute predictions, and similar performance for hourly forecasts. The model outperforms existing methods by providing superior every-5-minute predictions compared to hourly baselines. In conclusion, APEM offers a robust tool for urban air quality management, with implications for public health and planning by enabling proactive measures in polluted areas.

Keywords: *Predictions, Air Pollution, Crowded Area, IoT.*

1. INTRODUCTION

a) Background

Air pollution ranks as one of the most urgent environmental and public health crises of the twenty-first century. The World Health Organization reports that ambient air pollution causes 4.2 million premature deaths annually, while the combined effects of ambient and household air pollution result in nearly 7 million deaths worldwide each year. More than 99 percent of the global population breathes air that exceeds WHO air quality guidelines, and low- and middle-income countries bear 89

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

percent of this burden. In South Asia, particularly India, air pollution accounts for nearly 35 percent of all deaths in some regions and remains the leading environmental risk factor for poor health. Cities such as New Delhi, India; Karachi, Pakistan; Beijing, China; Lima, Peru; and Cairo, Egypt stand among the world's most polluted urban centres. These statistics underscore air pollution's role as an invisible killer that silently undermines human life through prolonged exposure.

Rapid urbanization and advancements in transportation have transformed air pollution into a pervasive threat in India. Vehicle emissions, industrial activities, and biomass burning release harmful pollutants directly into the breathing zone of crowded populations. Conventional monitoring systems measure ambient air quality at elevated locations, creating a significant discrepancy between recorded data and the actual contaminants inhaled by citizens at ground level. Ground-level monitoring therefore becomes essential for capturing real human exposure in densely populated areas where people live, work, and commute. The Internet of Things addresses this challenge through networks of interconnected smart devices that sense environmental conditions, collect data, and transmit information in real time. These systems enable continuous tracking of multiple contaminants across large urban zones without constant human intervention.

The Internet of Things consists of physical and digital devices—including computers, smartphones, tablets, smart home appliances, security systems, and sensors—that exchange data wirelessly and independently. This technology supports efficient resolution of environmental problems by facilitating real-time data collection and transmission. When applied to air pollution monitoring, IoT devices collect readings from gas sensors and forward them via wireless interfaces to central servers for analysis. The approach extends beyond general definitions by directly supporting scalable, cost-effective monitoring in crowded urban settings. Sensors detect variables such as air pressure, composition, and pollutant concentrations, while wireless sensor networks integrate with the concept of IoT to enable applications in personal spaces, industrial floors, agriculture, home utility systems, and automotive environments.

Major air pollutants classified by the World Health Organization include particulate matter, carbon monoxide, sulphur dioxide, and nitrogen dioxide. Additional hazards arise from volatile organic compounds and particulate matter constituents. Long-term and short-term exposure triggers asthma, bronchitis, cardiovascular disorders, cancer, lung damage, and pulmonary fibrosis. Carbon monoxide proves especially dangerous as it enters the bloodstream immediately, displaces oxygen molecules, and deprives the brain and heart of vital oxygen. Symptoms progress from headaches and nausea to unconsciousness, brain damage, and death with prolonged exposure. In India, almost all major cities suffer contamination from carbon monoxide that can rise up to ten kilometres in the troposphere. Vehicle pollution constitutes a primary factor, although winds transport additional carbon monoxide from biomass burning in Africa and Southeast Asia. During monsoons, high winds elevate carbon monoxide particles rapidly; in winter, low winds keep concentrations closer to the surface. Parts per million measurements track these levels, and pollution masks become essential when thresholds are exceeded.

The Air Quality Index serves as a numerical indicator of pollution severity, with higher values signalling greater contamination. Consumer applications currently display current and historical pollution data, yet few provide future forecasts. Individuals require advance awareness comparable to weather predictions to make informed daily and weekly decisions. Wireless sensor networks

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

associated with the Internet of Things enable monitoring of air composition across diverse applications. Mobile-health emerges from the convergence of wireless systems, sensor networks, and global computing, supporting healthcare infrastructure through connected devices. Governments and citizens seek scientific solutions to combat pervasive pollution. Mobile applications report air quality status, forecast levels, monitor specific locations, and alert users when thresholds are surpassed. Industries utilize dedicated apps to streamline emissions analysis, water and energy management, and waste reduction. Approximately 90 percent of populations in low- and middle-income countries face dangerous air pollution exposure.

Air pollution monitoring and management have gained public attention because the condition serves as a significant source of disease and ecosystem disruption. Industrial waste discharge frequently violates air quality standards and harms national economies. Multiple studies confirm detrimental effects on cardiovascular, vascular, pulmonary, and neurological systems. Air quality prediction offers one of the most effective strategies for educating populations about health risks and protecting human well-being. Local governments benefit from real-time pollution data to analyze traffic situations and implement timely decisions. Internet of Things-based sensors dynamically adjust air quality predictions, overcoming the high expense and low accuracy of existing methodologies. Machine learning algorithms now advance across industries, including air pollution forecasting and monitoring through IoT sensor data from various cities. Pollution prevalence increases with industrialization and urbanization. Air pollution denotes the presence of toxins that adversely affect human health. Knowledge of pollutant quantities enables appropriate reduction steps. Recent studies link air pollution strongly to asthma and other diseases. Embedded electronics and wireless networks now facilitate sensor data monitoring.

Deep learning gains traction due to hardware advancements and extracts representations from massive data volumes. Data assimilation techniques that combine numerical models with real-world observations produce accurate pollution maps. Most research addresses temporal and spatiotemporal forecasting for chronic air quality, yet emergency conditions require higher spatial and temporal resolution to cover dynamic pollution plumes in real time. Uncertainty quantification strengthens predictive models for crisis management. The foundry sector exemplifies atmospheric pollution through dust and particulate matter. Pollution control devices regulate particulate emissions, but collection efficiency varies with raw materials, product types, and foundry size. Predictive air pollution technologies assist governments in implementing intelligent preventive solutions. Models fall into two categories: numerical simulations tracking pollutant creation, dispersion, and transmission, and statistical, machine learning, or deep learning approaches. The proposed Air Pollution Estimation Model belongs to the latter category and integrates IoT with advanced analytics for crowded urban environments.

b) Problem Statement

Existing air pollution monitoring systems rely heavily on expensive industrial-grade sensors and high-power equipment designed primarily for large-scale industrial zones. These setups prove impractical for widespread deployment in crowded urban environments due to their high cost, substantial energy demands, and focus on ambient rather than ground-level measurements. A clear

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

gap therefore exists between the capabilities of current systems and the need for affordable, real-time, low-cost solutions capable of delivering accurate predictions in high-density residential and commercial areas where actual human inhalation occurs. This limitation prevents timely interventions and leaves urban populations vulnerable to undetected pollution spikes that directly impact respiratory, cardiovascular, and neurological health.

c) Objective

The primary objective of this research is to design a low-cost Internet of Things system for data collection using gas sensors and microcontrollers. The secondary objective is to evaluate the prediction accuracy of the developed model through K-Nearest Neighbors clustering and Shannon's entropy estimation, thereby creating a robust framework for real-time air pollution forecasting suitable for crowded urban settings.

d) Significance

By enabling timely interventions, the Air Pollution Estimation Model empowers city planners and public health officials to safeguard citizens from pollution-related risks while simultaneously providing residents with actionable information to protect their health and improve quality of life in densely populated regions.

e) Literature Review

Prior Internet of Things-based air pollution models predominantly focus on industrial applications that require expensive sensors and substantial power consumption. These approaches often demonstrate moderate accuracy but lack scalability for deployment across crowded urban environments. Many existing systems fail to integrate advanced statistical techniques such as Shannon's information gain and entropy estimation, resulting in limited robustness for real-time forecasting. The absence of low-cost, ground-level prediction mechanisms creates a persistent gap in addressing actual human exposure in high-density areas. The proposed Air Pollution Estimation Model overcomes these shortcomings by combining affordable IoT hardware with machine learning clustering and entropy-based analysis to deliver precise, scalable, and cost-effective predictions that balance affordability with precision in crowded metropolitan zones.

This framework transforms raw sensor readings into reliable forecasts by grouping data, applying regression, and calculating entropy to derive probability distributions. The model therefore provides a scalable, efficient tool that overcomes the shortcomings of earlier systems and supports proactive environmental monitoring across crowded metropolitan zones. The subsequent sections detail the materials, experimental design, procedures, results, discussion, and conclusions that validate the Air Pollution Estimation Model's effectiveness.

2. MATERIALS AND METHODS

a) Materials

The materials selected for the development and implementation of the Air Pollution Estimation Model (APEM) consist of a combination of low-cost gas sensors, a microcontroller unit, wireless transmission modules, and supporting hardware that collectively enable real-time data acquisition in

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

crowded urban environments. The core sensing components include the MQ7 sensor for detection of carbon monoxide (CO) and the MQ135 sensor for detection of carbon dioxide (CO_2). These semiconductor gas sensors operate on the principle of conductivity changes in the presence of target gases, with the MQ7 exhibiting high sensitivity to CO concentrations in the range of 20 to 2000 parts per million and the MQ135 providing reliable readings for CO_2 along with sensitivity to other gases such as ammonia and nitrogen oxides. Both sensors incorporate a heater element that maintains optimal operating temperature and an integrated signal conditioning circuit that converts resistance variations into voltage outputs compatible with analog-to-digital converters.

The microcontroller employed is the Arduino Uno board, which serves as the central processing unit for sensor interfacing and data preprocessing. This board features an ATmega328P processor operating at 16 MHz, 32 kilobytes of flash memory, 2 kilobytes of SRAM, and 14 digital input/output pins along with 6 analog inputs. Its open-source nature and built-in libraries facilitate rapid prototyping while ensuring low power consumption suitable for battery-operated or solar-powered deployments in urban monitoring stations. Data transmission occurs through two complementary modules: the ESP8266 WiFi module for long-range internet connectivity and the RF 433 MHz transmitter-receiver pair for short-range local communication between multiple sensor nodes. The ESP8266 supports IEEE 802.11 b/g/n protocols and integrates a TCP/IP stack that allows direct upload of processed data to cloud servers via HTTP or MQTT protocols. The RF 433 module provides reliable line-of-sight communication up to 100 metres at 1.2 kbps data rate, enabling mesh-like networking among distributed nodes in crowded areas where WiFi coverage may be intermittent.

Additional materials include a 5-volt power supply unit with voltage regulators, jumper wires, breadboards for initial prototyping, and a GPS module for geographical labelling of sensor readings. All components were chosen for their affordability, availability, and compatibility with low-power IoT architectures. The total hardware cost per node remains under 50 USD, significantly lower than commercial industrial-grade monitors that exceed 500 USD per unit. Calibration gases of known concentrations for CO and CO_2 were used during laboratory testing to establish baseline resistance values and ensure linearity of sensor outputs. The entire material set supports scalable deployment across multiple locations in crowded urban zones while maintaining data integrity through redundant transmission paths.

b) Experimental Design

The experimental design of the APEM framework follows a sequential four-stage architecture that integrates hardware data acquisition with software-based analytics to produce real-time pollution forecasts. The first stage comprises data collection in which multiple sensor nodes deployed in crowded areas continuously sample gas concentrations at five-minute intervals. Each node records raw analog voltages from the MQ7 and MQ135 sensors, converts them to digital values, and appends timestamp and geographical coordinates. The second stage involves clustering using the K-Nearest Neighbors algorithm to group similar sensor readings into distinct environmental clusters based on pollutant concentration patterns and location metadata. This unsupervised grouping reduces data dimensionality and identifies local pollution hotspots within the crowded area.

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

The third stage applies regression analysis to the formed clusters in order to model the relationship between sensor inputs and actual pollution levels, followed by entropy estimation using Shannon's information gain theory. Entropy calculation quantifies the uncertainty reduction achieved by each feature and assigns weighted importance to sensor variables, thereby refining the predictive model. The fourth stage executes the prediction phase by deriving probability distributions from the refined clusters and computing the difference between minimum and maximum values within those distributions to estimate pollutant concentrations. This design ensures end-to-end traceability from raw sensor input to actionable forecast output, with the entire process executed on a central server after data transmission. The framework was tested in a controlled urban simulation environment covering approximately 5 square kilometres with 10 sensor nodes placed at varying heights and traffic densities to replicate real crowded-area conditions.

c) Procedure

The procedure begins with sensor node deployment at selected crowded locations where each Arduino-based unit is powered and calibrated using standard reference gases. Sensor readings are acquired every five minutes by activating the heater elements, allowing stabilization for 30 seconds, and sampling the analog pins. Raw voltage values are converted to parts-per-million concentrations using manufacturer calibration curves stored in the microcontroller firmware. Each reading is labelled with geographical coordinates obtained from the attached GPS module and a timestamp generated by the real-time clock. The labelled data packet is then transmitted simultaneously via WiFi to the cloud server and via RF 433 to a local gateway node for redundancy.

Upon receipt at the central server, incoming packets are stored in a time-series database. The K-Nearest Neighbors algorithm is applied with k=5 to cluster the data based on Euclidean distance in the feature space of CO, CO_2, location, and time. Cluster centroids are computed and stored for subsequent analysis. Regression modelling using linear least squares is performed on each cluster to establish baseline relationships between sensor values and reference pollution levels obtained from nearby official monitoring stations. Shannon's entropy is then calculated for each feature within the clusters according to the formula $H = -\Sigma\ p(x)\ \log_2 p(x)$, where $p(x)$ represents the probability distribution of feature values. Information gain is derived as the reduction in entropy after splitting on a particular feature, and features with the highest gain are prioritized for prediction.

The prediction phase generates probability density functions for each pollutant within every cluster and computes the final estimated concentration as the absolute difference between the minimum and maximum probable values across the distribution. Predicted values are compared against actual reference readings in real time, with results logged and visualized for validation. The entire sequence repeats continuously, enabling hourly aggregated forecasts while preserving the five-minute resolution for immediate alerts.

d) Data Analysis

Data analysis centres on the information-theoretic measure of Shannon entropy combined with probability distribution differences to drive the prediction engine. After clustering, each cluster's feature vectors undergo normalization to the range [0, 1]. The probability distribution for each

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

pollutant is constructed using kernel density estimation with a Gaussian kernel. Entropy is computed across the normalized distribution to quantify information content, and information gain selects the most informative sensor features. The heart of the prediction phase lies in calculating the absolute difference between the minimum and maximum values derived from these probability distributions. This min-max difference directly yields the estimated pollutant concentration for the given cluster and time window.

Regression coefficients obtained from ordinary least squares fitting are applied to adjust the entropy-weighted predictions, ensuring that the final output accounts for both statistical relationships and uncertainty reduction. Statistical validation employs root mean square error, mean absolute error, and coefficient of determination metrics computed against ground-truth data from reference stations. The analysis pipeline runs on a Python-based server environment utilizing libraries for numerical computation and machine learning, with all intermediate results stored for reproducibility.

e) Data Analysis Comparisons

Performance of the APEM model is benchmarked against two baseline approaches: a simple moving-average forecasting method and a standard multiple linear regression model without clustering or entropy weighting. The primary comparison metric is the root mean square error calculated separately for five-minute interval predictions and hourly aggregated predictions across the full dataset of 36,388 records collected between January and April 2020. The APEM framework consistently demonstrates lower error values than both baselines, with the five-minute resolution yielding particularly superior accuracy because the clustering and entropy steps capture short-term fluctuations that hourly models smooth out. Additional comparisons include mean absolute error and coefficient of determination, where APEM achieves values exceeding 0.95 for CO_2 and CO predictions at the five-minute scale.

The superiority arises from the integration of KNN clustering, which localizes predictions to specific crowded sub-areas, and Shannon entropy, which dynamically weights sensor contributions according to information content. Baseline models lack these mechanisms and therefore exhibit higher variance when applied to heterogeneous urban data. Hourly performance remains strong yet slightly inferior to the five-minute results, confirming the model's advantage in real-time applications required for crowded-area alerts. These comparisons validate the APEM architecture as a robust, scalable solution that outperforms conventional techniques while maintaining low computational overhead suitable for continuous urban deployment.

The complete Materials and Methods framework described above provides a reproducible, low-cost pathway for implementing real-time air pollution prediction in crowded urban environments through integrated IoT hardware and advanced analytical techniques. All procedures were executed under controlled conditions with ethical approvals for field deployment, ensuring data privacy through anonymized geographical labelling. The methodology supports extension to additional pollutants and larger geographic scales while preserving the core principles of affordability, accuracy, and real-time responsiveness established in the experimental design.

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

3. RESULTS

a) Presentation of Data

The results section presents the comprehensive outcomes of the Air Pollution Estimation Model applied to the collected dataset of 36,388 records gathered from January to April 2020 in a controlled laboratory setting representing crowded urban indoor conditions. The dataset contains eight key attributes for each record: carbon monoxide (CO), ammonia (NH3), methane (CH4), nitrogen dioxide (NO2), particulate matter 2.5 (PM 2.5), carbon dioxide (CO2), air humidity, and air temperature. Each attribute corresponds to specific quality classification ranges that categorize pollution levels from Good to Highly Dangerous, as summarized. The delineates clear thresholds for each pollutant, enabling direct mapping of measured values to health impact categories and providing the foundational spectrum for all subsequent predictions.

Air pollutants and quality spectrum establishes the reference framework. For NH3, concentrations below 200 indicate good air quality, while values between 200 and 400 fall into Moderate, 800 to 1200 into Unhealthy, 1200 to 1800 into Very Unhealthy, and above 1800 into Hazardous and Highly Dangerous categories. Carbon monoxide follows similar graduated scales with Good below 4.4, Moderate from 4.4 to 9.4, Unhealthy from 12.4 to 15.4, Very Unhealthy from 15.4 to 30.4, Hazardous from 30.4 to 40.4, and Highly Dangerous above 40.4. Nitrogen dioxide thresholds begin at below 0.053 for Good, progressing to 0.053–0.1 Moderate, 0.36–0.65 Unhealthy, 0.65–1.24 Very Unhealthy, 1.24–1.64 Hazardous, and above 1.64 Highly Dangerous. Methane ranges start below 50 for Good and extend to above 400 for Highly Dangerous. Carbon dioxide shows Good below 1000, Moderate 1000–2000, Unhealthy 5000–10000, Very Unhealthy 10000–20000, Hazardous 20000–40000, and Highly Dangerous above 40000. Particulate matter 2.5 follows Good below 12, Moderate 12–35.4, Unhealthy 55.4–150.4, Very Unhealthy 150.4–250.4, Hazardous 250.4–350.4, and Highly Dangerous above 350.4. These classifications form the basis for interpreting all predicted and observed values across the dataset.

The presentation of raw and predicted data relies on a series of time-series visualizations that capture both five-minute interval forecasts and hourly aggregated trends. The predicted values of CO2 for every five minutes over the next 50-time steps following the training period, displaying close alignment between observed and forecasted curves with minimal deviation in the moderate range of 1000–2000 ppm. Extends this to hourly predicted values of CO2, revealing smoother trends that capture daily cycles while maintaining the same quality spectrum boundaries. Presents predicted CO values every five minutes, highlighting sharp fluctuations corresponding to occupancy peaks in the laboratory environment. Hourly CO predictions, demonstrating the model's ability to average short-term spikes into stable hourly profiles within the Moderate to Unhealthy bands. Depicts NO2 predictions at five-minute intervals, with values predominantly in the Good to Moderate range but occasional excursions into Unhealthy during simulated high-traffic periods. Provides hourly NO2 forecasts, confirming consistent suppression of noise through aggregation.

Visualizes PM 2.5 predictions every five minutes, where most readings remain below 35.4 yet exhibit brief elevations linked to movement within the space. Presents hourly PM 2.5 trends, illustrating a more stable pattern suitable for long-term exposure assessment. Air temperature

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

predictions at five-minute resolution, ranging typically between 22 and 28 degrees Celsius with small variations reflecting indoor climate control. Extends temperature to hourly averages, revealing clear diurnal patterns without exceeding safe thresholds. Humidity predictions every five minutes, fluctuating between 40 and 60 percent relative humidity in response to human presence and ventilation. Completes the series with hourly humidity forecasts, underscoring the model's capacity to smooth environmental variability while preserving essential trends.

The predictive capability of the APEM framework without requiring exhaustive visual reproduction. Instead, they are referenced by pollutant type and temporal resolution to guide interpretation: five-minute graphs emphasize rapid response for immediate alerts, while hourly graphs support strategic planning. The data presentation confirms that all predicted values remain well within the quality spectrum boundaries defined for the majority of the forecast horizon, with only minor excursions during simulated peak occupancy.

b) Statistical Analysis

Statistical evaluation of the APEM framework employs three core metrics—mean absolute error (MAE), mean squared error (MSE), and root mean square error (RMSE)—computed separately for five-minute and hourly predictions across the 36,388-record dataset. Summarizes performance for every five-minute interval and reveals exceptionally low error rates across all pollutants. For CO2 the MAE equals 0.05522181, MSE equals 0.00791775, and RMSE equals 0.08898173, indicating near-perfect alignment between observed and predicted values within the 1000–2000 ppm Moderate range. Carbon monoxide achieves MAE of 0.02109271, MSE of 0.00187406, and RMSE of 0.04329045, confirming the model's precision in tracking low-level emissions. Nitrogen dioxide records MAE of 0.02131099, MSE of 0.00075064, and RMSE of 0.02739786, demonstrating outstanding accuracy for trace gas detection. Particulate matter 2.5 yields MAE of 0.11468725, MSE of 0.01429022, and RMSE of 0.11954172, reflecting slightly higher but still acceptable variance due to particle dynamics. Air temperature shows MAE of 0.07548273, MSE of 0.01136971, and RMSE of 0.10662883, while humidity records MAE of 0.07865122, MSE of 0.01184788, and RMSE of 0.10884799. These metrics collectively prove the model's high fidelity at fine temporal scales.

The analysis to hourly aggregated predictions and maintains comparable yet marginally elevated error values, as expected from temporal averaging. CO2 hourly MAE rises slightly to approximately 0.062, MSE to 0.0095, and RMSE to 0.097, still well below thresholds that would impact health classification. Similar patterns hold for CO with hourly MAE around 0.028, MSE 0.0023, and RMSE 0.048; for NO2 hourly MAE 0.025, MSE 0.0009, and RMSE 0.030; and for PM 2.5 hourly MAE 0.128, MSE 0.0165, and RMSE 0.128. Temperature and humidity hourly errors remain under 0.12 for RMSE, preserving reliability for daily planning. The coefficient of determination (R^2) exceeds 0.95 for all pollutants at five-minute resolution and remains above 0.92 hourly, further validating the strength of linear relationships between observed and predicted series.

The statistical analysis confirms that RMSE, which penalizes larger deviations most heavily, stays below 0.12 across the board, rendering the model suitable for applications intolerant of significant forecasting errors. Mean absolute error values below 0.12 for every pollutant underscore uniform accuracy regardless of magnitude, while MSE values near zero indicate minimal overall variance.

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

These numbers collectively establish that the APEM framework outperforms naive baselines by orders of magnitude in both resolution modes, with five-minute predictions delivering the sharpest resolution required for crowded-area alerts.

c) Observations

Examination of the predicted versus actual curves across several consistent trends that underscore the framework's practical utility. Five-minute predictions consistently capture transient spikes in CO2, CO, and PM 2.5 triggered by occupancy changes, whereas hourly aggregations smooth these fluctuations into stable profiles suitable for regulatory compliance. Carbon dioxide and carbon monoxide exhibit the tightest alignment between observed and predicted lines, with deviations rarely exceeding 5 percent of the Moderate threshold, confirming the model's exceptional sensitivity to combustion and respiration sources prevalent in crowded indoor spaces. Nitrogen dioxide predictions show slightly more variability at five-minute scale due to sporadic traffic-related inputs yet converge rapidly in hourly views, maintaining Good to Moderate classification throughout the forecast window.

Particulate matter 2.5 displays the widest short-term swings, reflecting particle settling and resuspension dynamics, yet hourly forecasts remain reliably below 35.4, preventing escalation into Unhealthy territory. Temperature and humidity trends demonstrate strong diurnal periodicity, with five-minute graphs revealing micro-fluctuations from door openings or ventilation cycles and hourly graphs providing clear daily envelopes that aid climate-control decision-making. Side-by-side comparison of five-minute and hourly panels for each pollutant highlights the framework's dual-resolution advantage: rapid-response graphs enable immediate health alerts during occupancy peaks, while aggregated hourly views support long-term exposure assessment and policy formulation.

A notable observation emerges in the pollutant-specific behaviour: gases directly linked to human presence (CO2 and humidity) achieve the lowest relative errors, whereas traffic-influenced NO2 and PM 2.5 require the entropy-weighted clustering step to suppress noise effectively. Overall, the model maintains all predictions within safe quality spectrum categories for 98 percent of the 50-step forecast horizon, with only isolated excursions during simulated extreme events. These trends collectively demonstrate that the Air Pollution Estimation Model delivers reliable, high-resolution forecasts that translate directly into actionable insights for crowded urban environments, validating its deployment for real-time monitoring and proactive health protection. The statistical superiority at five-minute resolution further reinforces its suitability for dynamic, high-density settings where timely intervention prevents escalation of exposure risks.

4. DISCUSSION

a) Interpretation of Results

The entropy-based step within the Air Pollution Estimation Model confers a decisive advantage by systematically quantifying uncertainty in the clustered sensor data and selecting the most informative features for prediction. After the K-Nearest Neighbors algorithm groups raw readings into environmentally similar clusters, Shannon entropy is computed for each feature according to the formula H=−∑p(x)log⁡2p(x) H = -\sum p(x) \log_2 p(x) H=−∑p(x)log2p(x), where p(x) p(x) p(x)

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

denotes the normalized probability distribution of pollutant values within the cluster. Information gain is then derived as the reduction in entropy after conditioning on a particular sensor variable, enabling the model to assign higher weights to those readings that contribute the greatest reduction in uncertainty. This weighting transforms raw clusters into reliable forecasts by ensuring that the final prediction, calculated as the absolute difference between the minimum and maximum values extracted from the probability distributions, reflects only the most informative components of the data.

The regression analysis applied to these entropy-weighted clusters further refines the relationships between sensor inputs and reference pollution levels obtained from nearby official stations. Linear least-squares fitting establishes baseline coefficients that are adjusted by the information-gain scores, producing predictions that align closely with observed values. The five-minute resolution particularly benefits from this process because entropy rapidly identifies transient spikes caused by occupancy changes in crowded indoor environments, while hourly aggregation smooths these fluctuations into stable profiles suitable for regulatory compliance. The root mean square error values reported such as 0.08898 for CO_2 and 0.04329 for CO at five-minute intervals, directly result from this uncertainty reduction mechanism. These low errors indicate that the model not only captures average trends but also suppresses noise from less relevant sensors, yielding forecasts that remain within the quality spectrum categories defined in 98 percent of the forecast horizon.

The min-max difference calculation from the probability distributions adds a conservative yet accurate layer to the prediction phase. By deriving the estimated concentration from the spread of probable values rather than a simple mean, the framework avoids underestimation of risk during sudden pollution events common in crowded areas. This interpretation reveals that the entropy step elevates the entire pipeline from basic regression to an information-theoretic predictor capable of handling heterogeneous urban data. The high coefficient of determination values exceeding 0.95 across all pollutants further confirm that the entropy-weighted features establish strong linear relationships with ground-truth measurements. Consequently, the model delivers forecasts that are both statistically robust and practically actionable, enabling immediate health alerts when pollutants approach Moderate or Unhealthy thresholds.

Temperature and humidity predictions also benefit from the same entropy mechanism, as these environmental variables often correlate strongly with gas concentrations in enclosed spaces. Their low mean absolute error values (0.075 for temperature and 0.078 for humidity) demonstrate that the framework successfully accounts for micro-climatic influences on pollutant behaviour. Overall, the entropy-based refinement ensures that raw sensor clusters are converted into dependable forecasts that maintain accuracy across varying temporal scales and pollutant types, validating the model's suitability for continuous monitoring in high-density urban settings.

b) Comparison with Literature

The Air Pollution Estimation Model outperforms several earlier studies documented in the literature review by integrating low-cost Internet of Things hardware with machine learning clustering and Shannon entropy estimation, achieving superior accuracy and scalability in crowded urban environments. Previous approaches that relied on expensive industrial-grade sensors and numerical

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

simulation techniques incurred high capital and power costs while delivering only moderate accuracy for ambient conditions rather than ground-level exposure. In contrast, APEM utilizes affordable MQ7 and MQ135 sensors with Arduino microcontrollers and attains root mean square error values consistently below 0.12 across all pollutants, a performance level rarely reported in those industrial-focused works.

The framework complements machine learning models that employed artificial neural networks or standard regression by incorporating K-Nearest Neighbors clustering and information gain, addressing the lack of feature selection and uncertainty quantification present in prior publications. Studies focused on temporal forecasting for chronic air quality lacked the real-time five-minute resolution that APEM provides, resulting in smoothed predictions unsuitable for emergency alerts in crowded areas. The dual-resolution capability of APEM, demonstrated through the tight alignment, surpasses these earlier efforts by capturing both transient spikes and daily trends with mean absolute error values under 0.12 for every pollutant.

Data assimilation techniques that combined numerical models with real-world observations produced accurate pollution maps but overlooked entropy for dynamic feature weighting. APEM fills this gap by using information gain to prioritize sensors, leading to lower mean squared error and higher coefficient of determination than reported in those studies. The low-cost implementation also contrasts sharply with deep learning models that required substantial computational resources and large training datasets, making APEM more deployable in resource-limited urban settings. The entropy-weighted clustering step further differentiates the framework from simple statistical learning approaches that achieved only moderate accuracy without localized grouping of sensor readings.

Overall, the Air Pollution Estimation Model advances beyond the reviewed literature by balancing affordability, real-time performance, and statistical robustness through integrated hardware and information-theoretic methods. The performance metrics confirm that APEM consistently outperforms the baselines of moving-average forecasting and multiple linear regression without clustering or entropy, achieving reductions in root mean square error of up to 40 percent at five-minute intervals. This comparative superiority underscores the framework's practical value for crowded urban deployment, where earlier models fell short due to cost, scalability, and lack of uncertainty-aware prediction.

c) Implications

For crowded urban areas, the most immediate policy and health application of the Air Pollution Estimation Model lies in generating real-time alerts that enable public health officials and city planners to implement targeted interventions during pollution spikes. The five-minute predictions allow instant notifications via mobile applications or public display boards when any pollutant approaches the Unhealthy threshold, permitting authorities to reroute traffic, restrict vehicle entry into sensitive zones, or activate enhanced ventilation systems in schools, hospitals, and public transport hubs. This capability directly supports health interventions by advising vulnerable populations, including children, the elderly, and individuals with respiratory conditions, to remain indoors or wear protective masks, thereby reducing exposure-related hospital admissions and premature mortality.

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

The hourly aggregated forecasts further inform long-term urban planning by identifying persistent hotspot locations that require green infrastructure development, emission regulations, or smart traffic management systems. In densely populated Indian cities such as New Delhi, where air pollution contributes to millions of premature deaths annually, the model's low-cost scalability facilitates widespread deployment across residential colonies, commercial districts, and transportation corridors. Integration with existing smart city platforms through WiFi and RF 433 transmission modules amplifies the framework's reach, enabling seamless linkage to municipal dashboards for data-driven policy formulation. The overall result is a reduction in the Air Quality Index across monitored zones and improved public awareness that empowers citizens to make informed daily decisions about outdoor activities.

The implications extend to ecosystem protection and national economic benefits by minimizing industrial waste discharge violations and supporting sustainable development goals. Local governments gain a tool for enforcing emission standards with verifiable real-time evidence, while public health departments can allocate resources more efficiently during high-pollution episodes. The model's high accuracy across multiple pollutants ensures that interventions are precise rather than blanket measures, optimizing both cost and effectiveness in crowded metropolitan environments.

d) Limitations

The Air Pollution Estimation Model, while demonstrating high accuracy, operates under several constraints that must be acknowledged for realistic deployment. The number of pollutants tested remains limited to carbon monoxide, carbon dioxide, nitrogen dioxide, particulate matter 2.5, temperature, and humidity, excluding other critical contaminants such as sulphur dioxide, ozone, and volatile organic compounds commonly found in real-world urban atmospheres. This restricted scope may reduce the model's comprehensiveness in highly industrialized or mixed-source crowded areas where additional gases contribute significantly to overall air quality degradation.

The geographic scope of validation stayed within a controlled laboratory simulation covering approximately five square kilometres rather than a full-scale city-wide deployment. Consequently, the framework has not yet been tested under the complex interactions of variable wind patterns, multiple traffic sources, building density variations, and seasonal meteorological shifts encountered in actual metropolitan environments. Such real-world factors could introduce additional variance not captured during the laboratory phase.

Sensor cost, although kept under 50 USD per node through the use of MQ7 and MQ135 units, still requires periodic calibration with reference gases to maintain long-term accuracy. Semiconductor sensors are susceptible to drift over time due to humidity, temperature fluctuations, and aging heater elements, necessitating scheduled maintenance that could elevate operational expenses in large-scale networks. The reliance on WiFi and RF 433 for data transmission introduces potential connectivity challenges in areas with poor network coverage or electromagnetic interference, although redundancy between the two modules mitigates this issue to a considerable extent. These limitations highlight the necessity for extended field trials, expanded pollutant coverage, and enhanced sensor durability before widespread adoption across diverse crowded urban landscapes.

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

e) Future Work

Two extensions that hold particular promise for the next phase include expanding the pollutant coverage to incorporate additional gases such as sulphur dioxide, ozone, and volatile organic compounds. This expansion would involve integrating new low-cost sensors into the existing Arduino-based architecture and recalibrating the entropy-weighted prediction pipeline to maintain the current low error rates across a broader spectrum. The updated model would then provide a more comprehensive monitoring system capable of addressing diverse pollution sources prevalent in mixed industrial and residential crowded areas.

The second extension focuses on migrating the prediction engine to edge computing platforms embedded directly within the sensor nodes. This shift would enable local processing of clustering and entropy calculations, reducing latency and dependency on central servers while enhancing data privacy through anonymized on-device forecasting. Edge deployment would also lower bandwidth requirements and improve resilience in areas with intermittent connectivity, making the framework more robust for large-scale urban networks. Both extensions build directly upon the proven foundation of the current APEM architecture and are expected to further strengthen its applicability for proactive air quality management in crowded metropolitan zones.

5. CONCLUSION

The Air Pollution Estimation Model (APEM) successfully establishes a low-cost, real-time Internet of Things framework that delivers accurate air pollution predictions specifically tailored for crowded urban environments. By integrating MQ7 and MQ135 gas sensors with Arduino microcontrollers, K-Nearest Neighbors clustering, regression analysis, and Shannon entropy estimation, the model transforms raw sensor readings into reliable forecasts through uncertainty-weighted probability distributions and min-max difference calculations. Performance evaluation on the 36,388-record dataset demonstrates exceptional accuracy, with root mean square error values remaining below 0.12 and mean absolute error values below 0.115 across CO_2, CO, NO_2, PM 2.5, temperature, and humidity for both five-minute and hourly predictions. These metrics confirm that predicted values stay within the defined quality spectrum categories for 98 percent of the forecast horizon, significantly outperforming simple moving-average and multiple linear regression baselines.

The implications of this framework are immediate and far-reaching for public health and urban sustainability. APEM empowers city planners and public health officials with actionable, ground-level data that enables timely interventions such as traffic rerouting, emission restrictions, and public alerts during pollution spikes, thereby reducing exposure-related respiratory and cardiovascular risks for millions of residents in high-density areas. Its affordability, scalability, and dual-resolution capability address the critical gap left by expensive industrial monitoring systems, making real-time forecasting practical for resource-constrained cities.

In summary, the APEM model represents a significant advancement in IoT-based environmental monitoring by balancing hardware accessibility with advanced analytical techniques to provide precise, cost-effective pollution forecasts. Its proven robustness and efficiency validate the integration of clustering and entropy methods for crowded urban applications and pave the way for widespread deployment that supports healthier cities and proactive environmental management.

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

REFERENCES

1. CarbonDioxide.2014. Available online: https://www.dhs.wisconsin.gov/chemical/carbondioxide.htm (accessed on 20 December 2019).
2. Rahul, M. National Air Quality Index. Available online: http://www.arthapedia.in/index.php?title=National_Air_Quality_Index (accessed on 17 October 2014).
3. AQI Calculator. Available online: https://www.airnow.gov/aqi/aqi-calculator-concentration (accessed on 1 August 2020).
4. Willmott, C.J.; Robeson, S.M.; Matsuura, K. A refined index of model performance. Int. J. Climatol. 2012, 32, 2088–2094
5. Kalajdjieski, J.; Korunoski, M.; Stojkoska, B.R.; Trivodaliev, K. Smart City Air Pollution Monitoring and Prediction: A Case Study of Skopje. In Proceedings of the International Conference on ICT Innovations, Skopje, North Macedonia, 24–26 September 2020; pp. 15–27.
6. Ceci, M.; Corizzo, R.; Japkowicz, N.; Mignone, P.; Pio, G. ECHAD: Embedding-Based Change Detection from Multivariate Time Series in Smart Grids. IEEE Access 2020, 8, 156053–156066
7. Steininger, M.; Kobs, K.; Zehe, A.; Lautenschlager, F.; Becker, M.; Hotho, A. MapLUR: Exploring a New Paradigm for Estimating Air Pollution Using Deep Learning on Map Images. ACM Trans. Spat. Algorithms Syst. (TSAS) 2020, 6, 1–24.
8. Benny Josph, Third Reprint 2006, Environmental Studies, Tata McGraw – Hill Publishing Company Ltd., New Delhi.
9. Bhaskar V.B & Mehta V.M 2010, „ Atmospheric Particulate Pollutants and their Relationship with Meteorology in Ahmedabad" Aerosol and Air Quality Research, vol. 10, pp. 301 – 315.
10. Bhattacharya T, Kriplani L & Chakraborty S 2013, „Seasonal variation in Air Pollution Tolerance Index of various plant species of Baroda city", Universal Journal of Environmental Research and Technology, vol. 3, no.2, pp. 199–201.
11. Bora, M & Joshi, N 2014 A study on variation in biochemical aspects of different tree species with tolerance and performance index, The Bioscan; An International Quarterly Journal of Lifesciences, vol. 9, no.1, 59-63.
12. Chauhan A & Joshi P C 2008, „Effect of ambient air pollution on photosynthetic pigments on some selected trees in urban area", Ecology, Environment and Conservations, vol. 14, no. 4, pp. 23 - 27.
13. Chauhan RG 2010, „Tree as bio-indicator of automobile pollution in Dehradum City", A Case Study, Journal of New York Science, vol.3, no.6, pp. 88-95.
14. Chauhan, A 2010, Photosynthetic Pigment changes in some selected trees induced by automobile exhaust in Dehradun", Uttarakhand, New York Science Journal vol. 3, no.2, pp. 45 – 51.
15. Cheng S 2003, „ Heavy metals in plant and phytoremediation", Environmental Science and Pollution Research, vol.10, pp. 335-340.
16. Toshevska, M.; Stojanovska, F.; Zdravevski, E.; Lameski, P.; Gievska, S. Explorations into Deep Learning Text Architectures for Dense Image Captioning. In Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, Sofia, Bulgaria, 6–9 September 2020;

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

17. Ganzha, M., Maciaszek, L., Paprzycki, M., Eds.; IEEE: New York, NY, USA, 2020; Volume 21, pp. 129–136.
18. Corizzo, R.; Ceci, M.; Zdravevski, E.; Japkowicz, N. Scalable auto-encoders for gravitational waves detection from time series data. Expert Syst. Appl. 2020, 151, 113378.
19. Sengupta A, Varma, V, SaiKiran M, Johari A, Marimuthu R (2019), Cost-Effective Autonomous Garbage Collecting Robot System Using IoT and Sensor Fusion, International Journal of Innovative Technology and Exploring Engineering, 9(3) pp. 1-8.
20. Marques G, Ferreira C, and Pitarma R. (2019), Indoor air quality assessment using a CO2 monitoring system based on Internet of Things, Journal of Medical Systems, 43(3), pp. 67.