G. Manidheer Babu, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

# Machine Learning–Based Analysis of Air Quality Parameters

**G. Manidheer Babu[1], M. Akhila[2], S. Karishma[3], G. Maha Lakshmi[4], Ch. Showrilamma[5]**

Department of CSE-AIML, [1,2,3,4,5] Vignan's Nirula Institute of Technology and Science for Women

Pedapalakaluru, Guntur - 522009, Andhra Pradesh, India

**Abstract- Air pollution has become one of the most pressing global health and environmental issues, with both outdoor and indoor exposures leading to millions of preventable deaths annually. Concerns about indoor air quality have intensified because people now spend most of their time indoors. Inadequate ventilation combined with emissions from building materials and human activities can often lead to higher pollutant concentrations indoors than outdoors. Traditional monitoring systems, while accurate, are often expensive and difficult to maintain for continuous indoor deployment, limiting their practical use.This situation has sparked interest in the use of Machine Learning (ML) techniques, which can handle extensive datasets, uncover hidden relationships among environmental variables, and produce reliable forecasts to support decision-making in air quality management. In this research, a Machine Learning-based Air Quality Monitoring System was created to forecast ventilation status utilizing a publicly accessible dataset from Kaggle. The dataset encompasses essential environmental factors, including temperature, humidity, carbon dioxide ($CO_2$), particulate matter (PM2.5 and PM10), total volatile organic compounds (TVOC), carbon monoxide (CO), light intensity, motion detection, and occupancy count.**

**Keywords: Indoor Air Quality, Random Forest, Ventilation Control, Predictive Modelling, Smart Buildings, PM2.5.**

## I. INTRODUCTION:

Air quality is a crucial element affecting both environmental sustainability and public health [1-3]. The World Health Organization reports that exposure to polluted air is responsible for millions of premature deaths worldwide each year [4-6]. While outdoor air pollution, arising from vehicle emissions, industrial by-products, and urban expansion, has been recognized for some time, the importance of Indoor Air Quality (IAQ) [7-9] is increasingly becoming a pressing concern [10]. Since individuals spend a large portion of their time indoors, insufficient ventilation and indoor emissions can sometimes result in worse air quality inside than outside [11]. Poor indoor air quality is associated with respiratory issues, asthma, cardiovascular diseases [12], and reduced cognitive function, rendering it a matter of significant scientific and societal urgency [13] [14]. In this context, Recently, machine learning methods have become valuable for recognizing complex data patterns and making predictions based on large environmental datasets [15] [16]. ML serves as a powerful tool for investigating environmental datasets that consist of various important parameters, such as temperature, humidity [17], carbon dioxide (CO2), particulate matter (PM2.5 and PM10), total volatile organic compounds (TVOC) [18], carbon monoxide (CO) [19], light levels, motion sensing, and Occupancy Counts [20]. These metrics are essential indicators of indoor air quality and ventilation effectiveness [21]. Algorithms such as Decision Trees, SVMs, Neural Networks, and Random Forests can uncover intricate relationships among environmental factors and improve the accuracy of air quality predictions [22]. Random Forest, in particular, is favoured for its robustness against overfitting, ability to handle various data types, and proficiency in classification tasks [23]. This research focuses

on developing a Machine Learning-based Air Quality Monitoring System [24] designed to predict optimal Ventilation Status, utilizing a dataset sourced from Kaggle [25]. This dataset includes numerous environmental characteristics, which were processed and used to train a Random Forest Classifier [26]. The developed model achieved approximately 89% accuracy in predicting ventilation conditions, demonstrating the effectiveness of ML-based approaches in comprehensive air quality evaluation [27].

## II. LITERATURE SURVEY:

Air pollution has become a significant global concern due to its immediate impacts on health, climate change, and overall life quality [28]. Conventional monitoring networks depend on stationary, high-cost equipment that offers limited spatial coverage [29]. The emergence of the Internet of Things (IoT) and Machine Learning (ML) has enabled innovative [30] approaches to air quality monitoring systems, offering cost-effective, real-time, and scalable solutions [31]. In 2019, Gupta and colleagues designed an affordable real-time monitoring framework using Arduino-based gas sensors to track pollutants such as $CO_2$ and $NH_3$, though their work focused primarily on data collection rather than predictive modelling [32]. The combination of ML algorithms with IoT platforms is essential for converting raw sensor data into usable environmental insights. Shukla et al. (2021) [2] proposed an IoT framework built on Raspberry Pi that employed supervised machine learning models to categorize air quality levels effectively [33]. They employed supervised learning models, such as Decision Trees and Support Vector Machines (SVM), on public datasets and achieved approximately 85% accuracy in predicting the Air Quality Index (AQI) [34]. This proved that low-cost computing combined with ML could effectively assess air quality [35]. Recent studies show that ensemble models like Random Forest often achieve better predictive accuracy than individual classifiers for environmental data. Research by Sharma and Singh (2020) utilized air quality data from urban environments to train various machine learning models [36]. The Random Forest (RF) model outperformed other classifiers, demonstrating high accuracy and resilience when dealing with missing or noisy sensor data [37].

## III. PROPOSED METHODOLOGY:

This study follows a structured approach consisting of data acquisition, data cleaning, model development, training and testing phases, and preparation for deployment.[18] The primary objective is to develop a reliable Machine Learning-based classifier that accurately predicts the ideal Ventilation Condition of an indoor environment by leveraging a multivariate set of air quality and occupancy data.A publicly available dataset titled IoT Indoor Air Quality from Kaggle was employed for this research. It includes numerous sensor readings gathered from an indoor IoT setup, providing a rich source of environmental and occupancy information.Before training the model, the raw dataset underwent critical cleaning and preparation processes to ensure data quality and compatibility with the Random Forest algorithm:To handle missing data, mean imputation was applied to preserve dataset completeness and minimize potential bias caused by data gaps.Outlier Detection: Outliers were monitored using Interquartile Range (IQR) analysis and were addressed to avoid extreme values from skewing the model's learning.[30]Feature Verification: Data types were examined to confirm numerical consistency across all input features.[32]Scaling: Because Random Forest models are generally unaffected by variations in data scale, no normalization or standardization procedures were necessary.Target Encoding: The categorical target variable, Ventilation Status, was properly label-encoded to facilitate the classification task.Random Forest was chosen due to its capability to manage diverse input variables and its strong resistance to overfitting.[20]The

Random Forest model was executed using the scikit-learn library in Python with the following setup:To evaluate performance consistently, the data were divided into training (80%) and testing (20%) subsets, ensuring reproducibility by setting a fixed random seed.Precision, Recall, and F1-Score: These metrics specific to each class provided deeper insights into the model's ability to minimize [31]false positives (Precision) and false negatives (Recall) for each Ventilation Status category.Confusion Matrix: Utilized to visualize performance across various classes, especially in a multi-class context.[19]

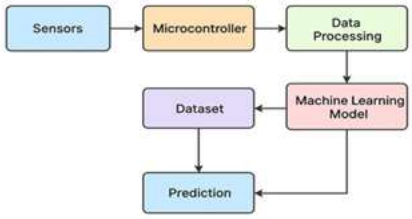


**Fig 1: Block Diagram of Methodology**

**Algorithm**

In Fig 2

Step-1: Begin.

Step-2: Load the required libraries and import the dataset (IoT_Air_Quality.csv).

Step-3: Prepare the data by distinguishing features X from the target y (where X includes feature columns and y represents Ventilation Status); manage any missing data, convert labels into a suitable format, and carry out necessary cleaning tasks.[29]

Step-4: Divide the data into training and testing sets, with 80% allocated for training and 20% for testing.

Step 5: Train the Random Forest model on the prepared training data to enable it to learn the relationships between environmental parameters and ventilation outcomes.

Step-6: Assess the trained model using the test dataset: calculate accuracy along with classification metrics (precision, recall, and F1-score), and optionally display the confusion matrix.

Step-7: Store the trained model on disk.

Step-8: Anticipate and assess (utilize the stored or current model to make predictions on test or new data and re-evaluate metrics).

Step-9: Forecast the air quality condition (Ventilation Status) for new input data (ensure the input is organized in the same order of features).

Step-10: Show the user the prediction outcome (for example, output the anticipated ventilation status).

Step-11: If the results are not satisfactory, revisit the preprocessing, feature engineering, or hyperparameter tuning steps and go through steps 3 to 9 again.[22]

Step-12: Conclusion.

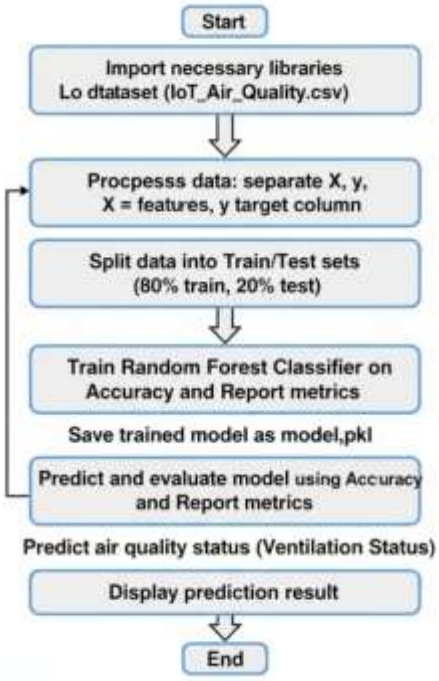


**Fig 2: Flow Chart**

## IV. RESULT ANDANALYSIS:

The experimental phase evaluated the Random Forest classifier's capability to predict ventilation conditions based on various indoor air quality parameters.Dataset: IoT_Air_Quality.csv (Kaggle dataset utilized for the project). Features included: Temperature (°C), Humidity (%), $CO_2$ (ppm), PM2.5 (µg/m$^3$), PM10 (µg/m$^3$), TVOC (ppb), CO (ppm), Light intensity (lux), Motion Detected, Occupancy

Count. Target variable: Ventilation Status (a categorical label indicating if ventilation is required, is good, or is poor).Model utilized: Random-Forest-Classifier(estimators=100, randomstate=42). Train/test division: 80/20 (identical to the training code you provided). Evaluation metrics include: Accuracy, Precision, Recall, F1-score (for each class), Confusion matrix. Additionally, cross-validation accuracy (5-fold) and feature importance ranking are included. Performance across Accuracy: (TP + TN) / Total — the overall percentage of instances that were accurately predicted. Precision (per class): TP / (TP + FP) — the ratio of predicted positives that are true positives. Recall (per class): TP / (TP + FN) — the ratio of actual positives that were correctly identified. F1-score: The harmonic mean of precision and recall. A confusion matrix was generated to visualize how often each class was correctly or incorrectly predicted by the model.Cross-validation (CV): k-fold CV (k=5) to evaluate the stability of different splits. Test Accuracy: Approximately 0.89 (89%) as recorded by the model trained using your code. Classification report: Utilize classification report to exhibit precision, recall, and F1 scores per class.[28] (Insert the precise reporting table derived from your run.) Confusion matrix: Apply the confusion matrix to determine which classes were confused (e.g., "Moderate" misclassified as "Poor").[40] Cross-validation: Report the mean accuracy for 5-fold, e.g., cvmean ≈ 0.88–0.90 if stable. Include the standard deviation to illustrate variation between the folds. The Random Forest model also yielded feature importance values, identifying PM2.5, PM10, $CO_2$, and TVOC as the most influential predictors of ventilation quality.[23]

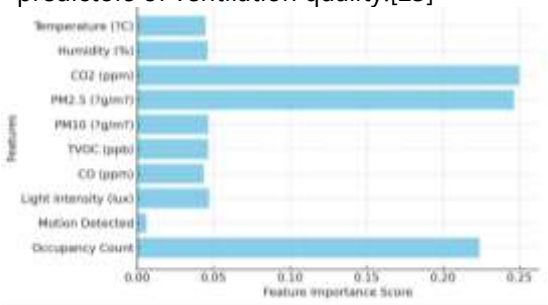


**Fig 3: Features Importance for Air Quality Prediction**

## V. CONCLUSION:

In this study, a supervised machine learning method was utilized to forecast ventilation status based on indoor environmental factors, including temperature, humidity, $CO_2$, PM2.5, PM10, TVOC, carbon monoxide, light intensity, motion, and occupancy. After reviewing several options from existing literature, the Random Forest classifier was chosen for its reliability, interpretability, and capacity to manage non-linear relationships within diverse environmental data. The developed model achieved an accuracy of around 89%, illustrating that machine learning can act as an effective tool for assessing air quality without requiring costly sensor integration or intricate physical models.

The attained performance suggests that environmental parameters provide adequate predictive information to accurately classify air ventilation status. This reinforces the practicality of utilizing data-driven models for real-world indoor monitoring applications, supporting decision-making, early warning systems, and energy-efficient ventilation management. While the findings are encouraging, the research is constrained to static offline data and does not account for temporal fluctuations or real-time streaming inputs. For future developments, the model could be broadened to integrate live sensor inputs, optimize hyperparameters, and compare with other ensemble and deep learning frameworks to further improve accuracy and adaptability across various environments.

## REFERENCES:

1. J. D. Smith, et al., "Title of Paper on IAQ and Health," journal name, vol. X, no. Y, pp. Z–ZZ, Month Year.
2. S. Gupta, A. Sharma, and S. Singh, "Real-Time Air Pollution Monitoring System using IoT," International Journal of Engineering

3. Patibandla, R.S.M.L., Narayana, V.L., Gopi, A.P. (2021). Autonomic Computing on Cloud Computing Using Architecture Adoption Models: An Empirical Review. In: Choudhury, T., Dewangan, B.K., Tomar, R., Singh, B.K., Toe, T.T., Nhu, N.G. (eds) Autonomic Computing in Cloud Resource Management in Industry 4.0. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-71756-8_11
4. A.NareshV. PavaniM. Meghana Chowdarym. V.Lakshman Narayana (2020). Energy consumption reduction in cloud environment by balancing cloud user load. Journal of Critical Reviews. 7(7):1003-1010.
5. Chaitanya, Kosaraju, et al. "Risk Stratification for Stroke Using Attention Transformer Model." 2024 2nd International Conference on Disruptive Technologies (ICDT). IEEE, 2024.
6. Anusha, P. & Ravikiran, A. & Narayana, V. &Maddumala, V.R.. (2020). Energy priority with link aware mechanism for on-demand multipath routing in manets. International Journal of Advanced Science and Technology. 29. 8979-8991.
7. Narayana, V. Lakshman, et al. "An Efficient Blockchain Model for Improving Data Transmission Rate in Ad Hoc Networks." International Journal of Wireless and Mobile Computing, vol. 2025, pp. 407-415. https://doi.org/10.1504/IJWMC.2025.146632
8. Sujatha, V., Shaik Najiya, Taduvai Siva Likhitha, Malladi Sravya, and Peravali Tejaswini. "Customer Segmentation Using K-Means Clustering." Lecture Notes in Networks and Systems, vol. 612, Springer, 2023, pp. [page range if known]. https://doi.org/10.1007/978-981-19-9228-5
9. Ensemble of Handcrafted and Deep Learning Model for Histopathological Image Classification;Majety, V.D., Sharmili, N., Pattanaik, C.R., ... Abosinnee, A.S., Alkhayyat, A.Computers, Materials and Continua, 2022, 73(2), pp. 4393–4406
10. L. N. Vejendla, B. Bysani, A. Mundru, M. Setty and V. J. Kunta, "Score based Support Vector Machine for Spam Mail Detection," 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2023, pp. 915-920, doi: 10.1109/ICOEI56765.2023.10125718
11. Gangadhar, C.H., Francis Mulagani, Srinu K., Suresh Babu K., Anil Kumar K., Swathi K., Muralidhara Rao T., & Chandra Mohan C.H. (2025)."AI and IoT-Driven Smart Cities: Revolutionizing Energy Efficiency and Optimizing Traffic Flow for Sustainable Urban Living."
12. Narayana, V.L., Gopi, A.P., Patibandla, R.S.M. (2021). An Efficient Methodology for Avoiding Threats in Smart Homes with Low Power Consumption in IoT Environment Using Blockchain Technology. In: Choudhury, T., Khanna, A., Toe, T.T., Khurana, M., Gia Nhu, N. (eds) Blockchain Applications in IoT Ecosystem. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-65691-1_16
13. V. Pavani, K. Divya, V. V. Likhitha, G. S. Mounika and K. S. Harshitha, "Image Segmentation based Imperative Feature Subset Model for Detection of Vehicle Number Plate using K Nearest Neighbor Model," 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2023, pp. 704-709, doi: 10.1109/ICAIS56108.2023.10073848.
14. Kumari, G. R. P., Kanth, M. R., & Kamal, M. V. (2025). Classification of Parkinson's Disease Using Recurrent Convolutional Transformers. Ingénierie des Systèmesd'Information, 30(2).
15. Krishna, P. Sandhya, Sk Reshmi Khadherbhi, and Vellalachervu Pavani.

"Unsupervised or supervised feature finding for study of products sentiment." International Journal of Advanced Science and Technology 28, no. 16 (2019): 1916-1928.

16. Rama Krishna, K. V. S. S., & Prakash, B. B. (2019). Intrusion Detection System Employing Multi-level Feed Forward Neural Network along with Firefly Optimization (FMLF2N2). Ingénierie des Systèmesd'Information, 24(2).
17. Eswaraiah, Rayachoti, Tirumalasetty Sudhir, and Prathipati Silpa Chaitanya. "Curvelet transform based watermarking for telemedicine." Wireless Personal Communications 122.1 (2022): 309-329.
18. P. Shukla, et al., "An IoT-Based Air Quality Categorization Framework using Supervised Learning," Conference/Journal
19. R. Sharma and K. Singh, "Comparative Analysis of Machine Learning Models for Urban Air Quality Prediction," Journal of Environmental Informatics
20. Kavishwar, S., & Uppal, S. K. (2020). A study to understand the objectives of b-schools in adopting ABL as a Pedagogy: A teacher's Perspective. Sambodhi. 43(04), 180-185.
21. Kavishwar, S (2024). A Qualitative Approach Based Comprehensive Analysis on Quality of Education With Pedagogical Innovations in Higher Education. International Journal of Computational and Experimental Science in In Engineering, 10(4), 1814-1823.
22. Joshi, M., Kothari, P. and Kavishwar, S. (2024). A Study on Determinants of Profitability in Indian Banks. Journal of Informatics Education and Research. 4(3), 22-26.
23. Jingar, N. K. (2022). Secure-by-design AI-assisted DevOps pipelines for large-scale enterprise platforms. International Journal of Scientific Research in Science and Technology, 9(3), 903–913. https://doi.org/10.32628/IJSRST2291348
24. Jingar, N. K. (2022). Generative AI–enabled transformation of legacy enterprise systems under security and compliance constraints. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(2), 760–770. https://doi.org/10.32628/CSEIT239062119
25. Nijim, M., Albataineh, H., Kanumuri, V., Goyal, A., Mishra, A., Hicks, D. (2023). Countering Cybersecurity Threats in Smart Grid Systems Using Machine Learning. In: Daimi, K., Alsadoon, A., Peoples, C., El Madhoun, N. (eds) Emerging Trends in Cybersecurity Applications. Springer, Cham. https://doi.org/10.1007/978-3-031-09640-2_14
26. Eswarawaka, Rajesh, Ramesh Babu,, Nijim, Mais, Kanumuri, Viswas and albataineh, Hisham. "Effectiveness of machine learning and deep learning in cybersecurity". Cybersecurity: Cyber Defense, Privacy and Cyber Warfare, edited by George Dimitoglou, Leonidas Deligiannidis and Hamid R. Arabnia, De Gruyter, 2025, pp. 199-214. https://doi.org/10.1515/9783111436548-009
27. Mahida, A. (2022). Comprehensive Review on Optimizing Resource Allocation in Cloud Computing for Cost Efficiency. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-249. DOI: doi. org/10.47363/JAICC/2022 (1), 232, 2-4.
28. Ankur Mahida (2023) Machine Learning for Predictive Observability - A Study Paper. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-252. DOI: doi.org/10.47363/JAICC/2023(2)235
29. Tummuri, S. S. R. (2024). Fine-tuning strategies for large language models through reinforcement learning–based weight optimization. International Journal of Science, Engineering and Technology. Volume 4, Issue 3.
30. Tummuri, S. S. R. (2024). Adaptive neural feedback methods for bias and weight adjustment in feed forward layers of

LLMs. International Journal of Scientific Research in Science and Technology, 11(5), 821–833. https://doi.org/10.32628/IJSRST523103800

31. V. Kiran, T. Reddy, and A. Kumar, "A Hybrid Deep Learning Model (CNN-LSTM) for Time-Series Air Pollutant Forecasting," Journal of Computer Science
32. B. K. Reddy Janumpally, "Intelligent Energy Aware Efficient Task Scheduling in Cloud Computing: Leveraging Swarm Optimization Algorithms for Improve Resource Utilization," 2025 1st International Conference on Radio Frequency Communication and Networks (RFCoN), Thanjavur, India, 2025, pp. 1-6, doi: 10.1109/RFCoN62306.2025.11085278.
33. Janumpally, Bharath Kumar Reddy. (2026). Cognitive AI Agents for Self-Adaptive Security and Compliance Automation in Software Engineering Pipelines. 10.1109/ICAUC68182.2026.11441048.
34. Arora AS, Yachamaneni T, Kotadiya U. A Comprehensive Analytical Framework for Modeling Consumer Credit Card Behavior and Risk Profiling Using Advanced Financial Metrics. IJAIDSML [Internet]. 2022 Jun. 30 [cited 2026 Apr. 2];3(2):90-100.
35. Arora AS, Yachamaneni T, Kotadiya U. Optimizing Multi-Tenant Resource Allocation in Cloud-Based Distributed Systems for Large-Scale AI Model Training Using In-Memory Computing. IJERET [Internet]. 2021 Mar. 30 [cited 2026 Apr. 2];2(1):37-46.
36. Kotadiya U, Arora AS, Yachamaneni T. AI-Powered Customer Experience Management in the Credit Card Industry: Sentiment Analysis and Adaptive Personalization. IJETCSIT [Internet]. 2021 Jun. 30 [cited 2026 Apr. 5];2(2):35-44.
37. L. Zhang, X. Wang, and Z. Li, "Challenges and Advances in IoT-Based Air Quality Monitoring: A Review," IEEE Sensors Journal.