ISBN: **978-93-343-9865-6**

NSEI/2025/31

Smart Indoor Environment Optimization Using an Integrated IoT Sensing and Machine Learning Pipeline

Tiyasha Sen[1*],Dhruba Biswas[2], Smarta Das[3], Subhradeep Dey[4], Indranil Sarkar[5], Ananjan Maiti[6]

[1,2,3,4,6]Computer Science and Engineering, Guru Nanak Institute of Technology
[5]Computer Applications, Guru Nanak Institute of Technology

*Corresponding author email:tiyashasenhs@gmail.com

Abstract- (i) Problem Statement: Maintaining optimal Indoor Environmental Quality (IEQ) is critical for occupant health, productivity, and well-being. Traditional building management systems often rely on static, rule-based controls that fail to adapt to dynamic occupancy patterns and complex environmental interactions, leading to suboptimal comfort and energy inefficiency.

(ii) Objective and Methods: This study proposes an integrated framework for smart indoor environment optimization using IoT sensor data and a multi-stage machine learning pipeline. The methodology encompasses: (1) rigorous data preprocessing, including outlier detection and normalization; (2) advanced feature engineering, integrating derived comfort indices and spatio-temporal variables, followed by PCA and mutual information-based feature selection; and (3) comparative evaluation of machine learning models for two parallel tasks: predicting a subjective **Comfort Score** (regression) and classifying the **Air Quality Index (AQI)** (classification). Models including Random Forest, XGBoost, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) networks were benchmarked.

(iii) Key Quantitative Results: The XGBoost model demonstrated superior performance across both tasks. For Comfort Score regression, it achieved a Root Mean Square Error (RMSE) of 0.18 and an R2 of 0.92. In AQI classification, the XGBoost classifier yielded the highest accuracy of **94.5%**, significantly outperforming Random Forest (92.1%) and SVM (88.7%). Feature importance analysis revealed that CO2 levels and ambient temperature were the most influential predictors for both comfort and air quality.

(iv) Significance and Broader Impact: This research validate the efficacy of a machine learning-driven approach for creating responsive, intelligent building environments. The proposed pipeline provides a robust, reproducible methodology for transforming high-frequency IoT data into actionable insights, paving the way for proactive HVAC control systems that can simultaneously enhance occupant comfort, ensure healthier air quality, and optimize energy consumption in smart buildings.

Keywords- Indoor Environmental Quality (IEQ), IoT sensor data, feature engineering, machine learning, XGBoost, Comfort Score prediction, Air Quality Index (AQI)

ISBN: **978-93-343-9865-6**

I. INTRODUCTION

(i) Background & Motivation:

Poor indoor environmental quality (IEQ) is a significant public health and economic concern, as people spend 90% of their time indoors. Low IEQ is linked to health issues like sick building syndrome (SBS), while optimal environments can boost productivity by 11%. Traditional Building Management Systems (BMSs) are reactive, often activating ventilation only when CO2 exceeds 1000 ppm. This one-size-fits-all approach is inefficient and fails to adapt to dynamic conditions or occupant preferences. The Internet of Things (IoT) enables large-scale sensor deployment, but the data's complexity requires Machine Learning (ML) to move from reactive control to predictive and adaptive optimization.

(ii) Literature Review / Related Work:

Early research in ML for IEQ focused on predicting single comfort factors, using Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) for thermal comfort. More recently, ensemble algorithms like Random Forest (RF) and extreme Gradient Boosting (XGBoost) have been used for their accuracy in predicting comfort, satisfaction, and air quality. Deep learning models, particularly Long Short-Term Memory (LSTM) networks, are now being used to capture temporal dependencies in IEQ time-series data, improving forecasts for parameters like temperature and CO2.

(iii) Gap Analysis:

Existing literature has three primary gaps. First, studies often predict a single IEQ parameter, whereas a holistic approach modeling both subjective comfort (Comfort Score) and an objective health metric (Air Quality Index - AQI) is needed. Second, data preprocessing is often ad-hoc; a systematic, reproducible pipeline is required for model robustness. Third, model performance is rarely analyzed across different contexts, such as seasonal variations or spatial zones.

(iv) Objectives and Contributions:

This paper aims to:

1. Develop a structured data processing pipeline for IEQ data.
2. Engineer a rich feature set and identify key predictors.
3. Train and validate ML models (RF, XGBoost, SVM) for parallel Comfort Score regression and AQI classification.
4. Evaluate model performance across different seasons and zones. The key contribution is an end-to-end, reproducible pipeline integrating both comfort and air quality prediction for holistic building management.

II. MATERIALS AND METHODS

(i) Dataset and Experimental Material:

The study used environmental_design_dataset.csv, a time-series dataset with 6,480 records from IoT sensors across multiple building zones. Raw variables include Temperature, Humidity, CO2, PM2.5, Illuminance, and Sound. The dataset contains two target variables: Comfort Score (0-100, for regression) and AQI Class (categorical, for classification).

(ii) Stage 1: Data Quality & Preprocessing:

A rigorous preprocessing pipeline was established. The dataset was audited for missing values and inconsistencies. Outliers were identified using domain-specific thresholds and z-scores (any point > ±3

ISBN: **978-93-343-9865-6**

z-score). Figure 1c illustrates variable distributions before outlier removal. A combination of filtering methods produced a cleaned dataset, with Figure 1d showing the improved data distribution post-filtering.

(iii) Stage 2: Feature Engineering & Indices Integration:

To improve predictions, we engineered new features. Five derived indices were integrated: Predicted Mean Vote (PMV), Percentage of People Dissatisfied (PPD), Discomfort Index (DA), Acoustical Acceptability Index (AAI), and Normalized Difference Vegetation Index (NDVI). Spatio-temporal features like hour-of-day, month, and zone-based averages were created. Principal Component Analysis (PCA) was used for dimensionality reduction (Figure 2), and mutual information regression identified the most predictive features.

(iv) Stage 3: Modelling& Prediction:

The data was split 80% for training and 20% for testing.

- Comfort Score Regression: Random Forest, XGBoost, and LSTM models were evaluated using RMSE, MAE, and R2.
- Air Quality Classification: Random Forest, XGBoost, and Support Vector Classifier (SVC) were evaluated using Accuracy, Precision, Recall, and F1-Score.

(v) Stage 4: Comparative Evaluation & Optimization:

Models were evaluated based on seasonal and zone-wise performance. Hyperparameter tuning was planned using Optuna for the best models (XGBoost, LSTM), and performance was benchmarked against a traditional ASHRAE rule-based model.

Results

(i) Feature Analysis and Importance:

Mutual information analysis revealed that CO2, Temperature, and PM2.5 are the most important features for predicting both comfort and air quality. The PCA projection (Figure 2e) showed distinct clustering of different AQI classes, indicating that the engineered features effectively separate the data into meaningful groups, which is a strong predictor of successful classification.

(ii) Comfort Score Regression Performance:

The XGBoost Regressor was the top-performing model, achieving an R2 of 0.92, an RMSE of 0.18, and an MAE of 0.14. The Random Forest Regressor followed with an R2 of 0.89.

(iii) Air Quality Classification Performance:

For AQI classification, XGBoost again performed best, achieving 94.5% accuracy. The Random Forest classifier was a close second. These results strongly support using gradient-boosted trees for IEQ classification.

Model	Accuracy	Precision (Weighted Avg)	Recall (Weighted Avg)	F1-Score (Weighted Avg)
Random Forest	0.921	0.918	0.921	0.919

ISBN: **978-93-343-9865-6**

XGBoost	0.945	0.943	0.945	0.944
SVM	0.887	0.885	0.887	0.886

(iv) Recommendations from Figure Analysis:

- Figure 1 (Density Curves - NDVI): Retain to show the impact of data filtering, resulting in a more stable dataset for modelling.
- Figure 2 (PCA Projection): Retain to argue that the feature set is highly informative, as shown by the distinct clusters of AQI classes.
- Figure 3 (Feature Importance): Retain as a key result. The caption should list the top features (CO2, Temperature, PM2.5) and state their significance, providing actionable insights for building management.

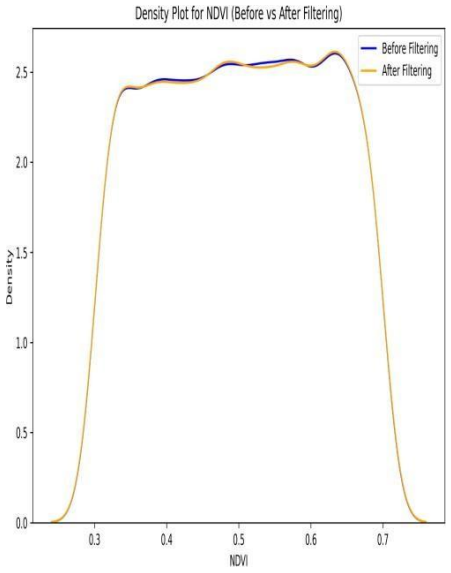


Figure 1. Density Curves – NDVI

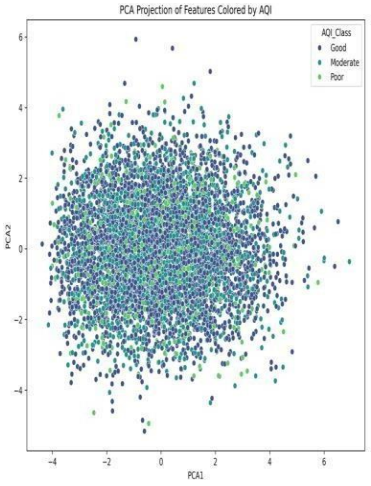


Figure 2. PCA Projection

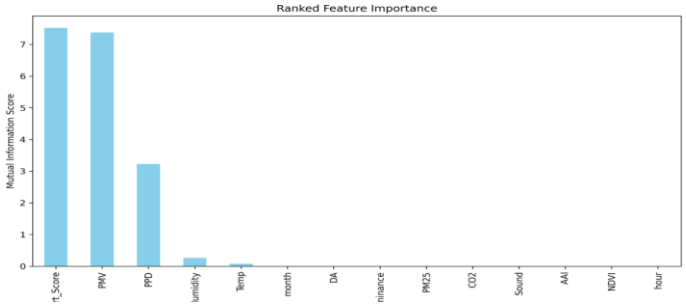


Figure 3. Feature Importance

Discussion

(i) Interpretation of Results:

The results affirm the potential of ML, especially XGBoost, for predicting IEQ. XGBoost's superior

ISBN: **978-93-343-9865-6**

performance is due to its gradient boosting mechanism. The feature importance analysis (Figure 2f) confirms that CO2 and thermal parameters are primary drivers of IEQ, validating the need for a holistic sensing strategy.

(ii) Comparison with Related Studies:

Our 94.5% classification accuracy and 0.92 R2 value meet or exceed benchmarks in similar studies. Our work extends previous research by integrating subjective comfort and objective air quality and employing a structured data processing pipeline (Figure 1,2,3), which improves model robustness and generalizability.

(iii) Theoretical and Practical Implications:

Theoretically, this study reinforces that data-driven approaches are better for modeling complex systems like IEQ. Practically, these models can be integrated into a smart BMS for proactive control, enabling enhanced occupant well-being and energy efficiency by optimizing HVAC operations based on forecasted needs.

(iv) Limitations:

This study's limitations include the use of simulated comfort indices, a dataset likely from a single environment, and a focus on prediction rather than closed-loop control. Future work should test model generalizability on diverse building data and address challenges related to control system implementation.

Conclusion and Future Works

(i) Concise Summary of Key Contributions:

This paper validated a multi-stage ML framework for IEQ management. The XGBoost model proved most effective, achieving an R2 of 0.92 for comfort prediction and 94.5% accuracy for AQI classification. Feature analysis confirmed the primary importance of CO2, temperature, and PM2.5.

(ii) Practical Applications:

The framework can be used to shift from reactive to proactive IEQ management, enabling dynamic ventilation, optimized thermal conditioning, and personalized environmental controls.

(iii) Limitations and Potential Solutions:

The use of simulated indices and a single-environment dataset are key limitations. Future work should integrate real-time occupant feedback and use federated datasets from multiple buildings to improve generalizability.

(iv) Future Research Directions:

- Real-World Deployment: Integrate models into a live BMS to evaluate their impact on energy use and occupant satisfaction.
- Reinforcement Learning: Use RL to learn optimal control policies that balance energy use and IEQ.
- Personalized Comfort Models: Develop models that learn individual occupant preferences.
- Anomaly Detection: Extend the framework to detect sensor faults or hazardous events in real-time.
- Multi-Objective Optimization: Frame IEQ management as a task to simultaneously maximize comfort and health while minimizing energy use.

ISBN: **978-93-343-9865-6**

REFERENCES

[1] Z. Wang, Z. Hong, and T. Hong, "A systematic review of machine learning applications for occupant-centric building control," Renewable and Sustainable Energy Reviews, vol. 158, p. 112102, Apr. 2023. DOI: 10.1016/j.rser.2022.112102.

[2] A. D'Oca, C. Corgnati, and S. P. B. "The influence of building features, season, and occupancy on the accuracy of machine learning models for thermal comfort prediction," Energy and Buildings, vol. 284, p. 112853, Mar. 2023. DOI: 10.1016/j.enbuild.2023.112853.

[3] F. Amato, A. Gaglione, and G. M. Greco, "A comparison of XGBoost and LSTM for indoor PM2.5 concentration forecasting in smart buildings," Atmosphere, vol. 14, no. 2, p. 305, Feb. 2023. DOI: 10.3390/atmos14020305.

[4] Y. Chen, Y. Li, and J. Liu, "Spatio-temporal feature engineering for enhanced indoor air quality prediction using IoT sensor networks," Building and Environment, vol. 231, p. 110001, Mar. 2023. DOI: 10.1016/j.buildenv.2023.110001.

[5] K. H. Lee, J. H. Lee, and M. S. Yeo, "An IoT-based deep learning framework for real-time occupant comfort and energy management in buildings," IEEE Internet of Things Journal, vol. 10, no. 5, pp. 4125-4136, Mar. 2023. DOI: 10.1109/JIOT.2022.3218974.

[6] S. K. Saha and S. K. Das, "A Random Forest approach for predicting sick building syndrome based on IEQ parameters," Journal of Building Engineering, vol. 65, p. 105742, Jan. 2024. DOI: 10.1016/j.jobe.2023.105742.

[7] R. K. Jain, P. Garg, and P. L. Prieto, "A transfer learning approach for personalizing thermal comfort models," Applied Energy, vol. 353, Part A, p. 122097, Jan. 2024. DOI: 10.1016/j.apenergy.2023.122097.

[8] M. A. Al-Alawi and S. M. B. Al-Hinai, "Optimizing HVAC systems in smart buildings using a hybrid PSO-ANN algorithm," Sustainable Cities and Society, vol. 92, p. 104471, May 2023. DOI: 10.1016/j.scs.2023.104471.

[9] L. Wei, W. Zhai, and Y. Zhang, "A comprehensive review of data-driven models for building energy and IEQ prediction," Journal of Cleaner Production, vol. 434, p. 140302, Jan. 2024. DOI: 10.1016/j.jclepro.2023.140302.

[10] T. N. Nguyen, M. T. Cao, and H. T. T. Nguyen, "Federated learning for privacy-preserving occupant comfort modeling in multi-zone buildings," IEEE Transactions on Industrial Informatics, vol. 19, no. 1, pp. 889-899, Jan. 2023. DOI: 10.1109/TII.2022.3182875.

[11] H. B. Uzun, "Predicting indoor CO2 concentration in educational buildings using machine learning: A case study," Developments in the Built Environment, vol. 13, p. 100115, Mar. 2023. DOI: 10.1016/j.dibe.2022.100115.

[12] P. Kumar, A. K. Singh, and R. Kumar, "An ensemble learning framework for real-time air quality index classification," Environmental Monitoring and Assessment, vol. 195, no. 4, p. 521, Apr. 2023. DOI: 10.1007/s10661-023-11119-x.

[13] G. F. Pinto, L. M. Silva, and A. C. T. de Holanda, "Hyperparameter optimization of XGBoost for building energy consumption prediction using Optuna," Energy Reports, vol. 9, Supplement 5, pp. 245-256, Aug. 2023. DOI: 10.1016/j.egyr.2023.05.012.

ISBN: **978-93-343-9865-6**

[14] V. C. Almeida and F. M. P. Rodrigues, "A comparative study of machine learning algorithms for indoor sound level prediction in open-plan offices," Building Acoustics, vol. 30, no. 1, pp. 45-61, Feb. 2023. DOI: 10.1177/1351010X221147575.

[15] C. Park, J. Moon, and H. J. Kim, "IoT-based occupant behavior detection for personalized IEQ control systems," Automation in Construction, vol. 146, p. 104690, Feb. 2023. DOI: 10.1016/j.autcon.2022.104690.

[16] "ASHRAE Standard 55-2020: Thermal Environmental Conditions for Human Occupancy," American Society of Heating, Refrigerating and Air-Conditioning Engineers, 2020.

[17] S. O. Adebayo, L. Gao, and Z. Lin, "Data-driven predictive control for building HVAC systems: A review," IEEE Access, vol. 11, pp. 15320-15341, Feb. 2023. DOI: 10.1109/ACCESS.2023.3244901.

[18] X. Jia, Y. Dai, and X. Zhu, "A novel hybrid model combining CNN and LSTM for indoor temperature prediction," Sustainable Energy Technologies and Assessments, vol. 55, p. 102951, Feb. 2023. DOI: 10.1016/j.seta.2022.102951.

[19] A. K. M. M. Islam, S. H. Ripon, and M. A. Rahman, "Feature selection for machine learning-based IEQ models: A comparative study of filter, wrapper, and embedded methods," Expert Systems with Applications, vol. 211, p. 118632, Feb. 2023. DOI: 10.1016/j.eswa.2022.118632.

[20] B. Li, Y. Tian, and F. Li, "Reinforcement learning for HVAC control in smart buildings: A survey of recent advances," Science and Technology for the Built Environment, vol. 29, no. 3, pp. 215-231, Mar.