Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 6, December 2026, pp. 5254~5264
ISSN: 2302-9285, DOI: 10.11591/eei.v15i6.11705

Optimized multi-class classification of PM2.5 in wooden houses at different altitudes

Muhamad Fuat Asnawi[1], Muhammad Alif Muwafiq Baihaqy[1], Adinda Septi Hendriani[2], Nulngafan[1], Hermawan[2]
[1]Department of Informatics Management, Faculty of Engineering and Computer Science, Universitas Sains Al-Qur'an, Wonosobo, Indonesia
[2]Department of Architecture, Faculty of Engineering and Computer Science, Universitas Sains Al-Qur'an, Wonosobo, Indonesia

Article Info

Article history:

Received Oct 24, 2025
Revised Sep 8, 2026
Accepted Sep 12, 2026

Keywords:

Air quality classification
CatBoost
Explainable artificial intelligence
PM2.5
Random forest
Shapley additive explanations
Support vector machine

ABSTRACT

This study investigates multiclass classification of PM2.5 concentrations in wooden houses located at different altitudes in Indonesia using optimized machine learning models. The dataset comprised approximately 430,000 observations from highland and lowland residential environments and included particulate, gaseous, meteorological, temporal, and location variables. Support vector machine (SVM), random forest (RF), and CatBoost were evaluated using an 80:20 stratified train-test split and five-fold cross-validation. SVM and RF were tuned using grid search, whereas CatBoost used randomized hyperparameter search with 25 candidate configurations. Optimization improved all models. RF achieved the best overall performance with 98.19% accuracy, 0.9763 macro F1-score, and 0.9655 matthews correlation coefficient (MCC), followed by CatBoost with 97.12% accuracy and SVM with 94.71% accuracy. The moderate class remained the most difficult because its feature patterns overlapped with adjacent categories. For CatBoost, optimization reduced test-set classification errors from 3,578 to 2,474 (30.9%). Shapley additive explanations (SHAP) analysis of the optimized CatBoost model using 5,000 independent test observations identified PM10 as the dominant predictor (mean |SHAP|=2.9566), followed by eCO2 and total volatile organic compounds (TVOC), while location provided a smaller but measurable contextual contribution (0.1696). The results show that ensemble models provide strong multiclass performance for heterogeneous residential PM2.5 data and that SHAP can clarify the environmental factors underlying model decisions.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Muhamad Fuat Asnawi
Department of Informatics Management, Faculty of Engineering and Computer Science
Universitas Sains Al-Qur'an
Andongsili, Mojotengah District, Wonosobo Regency, Central Java, 5635, Indonesia
Email: fuatasnawi@unsiq.ac.id

1. INTRODUCTION

Air pollution remains one of the most critical environmental and public health challenges worldwide. Among various pollutants, fine particulate matter (PM2.5) has received particular attention because of its ability to penetrate deep into the respiratory system and bloodstream, leading to cardiovascular disease, respiratory disorders, neurological impacts, and premature mortality. Accurate monitoring and classification of PM2.5 concentrations are therefore essential for supporting environmental policy, health protection, and early warning systems [1], [2].

Recent advances in machine learning have significantly improved air quality prediction and classification compared with traditional statistical approaches. Algorithms such as support vector machine (SVM), random forest (RF), gradient boosting, and deep learning models have demonstrated strong capability in capturing nonlinear relationships among pollutants, meteorological factors, and temporal patterns [3], [4]. In particular, ensemble learning methods have shown superior robustness when handling noisy environmental datasets, high-dimensional variables, and class imbalance problems commonly found in air quality applications [5], [6].

Although many previous studies have reported promising results, most PM2.5 research has focused on urban monitoring stations, metropolitan datasets, satellite observations, or regional forecasting systems [7], [8]. Comparatively limited attention has been given to housing-specific indoor or semi-indoor environments, where residents spend a substantial portion of their daily lives. Exposure patterns in residential buildings may differ considerably from urban outdoor stations due to ventilation behavior, building materials, household activities, and surrounding microclimatic conditions [9], [10].

This limitation becomes more significant in wooden houses located at different altitudes, where structural and environmental factors can substantially alter pollutant accumulation and dispersion. Highland houses are typically characterized by lower ambient temperatures, distinct wind circulation, and heating-related activities, while lowland houses may experience warmer conditions, denser surrounding emissions, and different humidity dynamics. These altitude-related microclimatic differences can influence PM2.5 concentration patterns and exposure risk, yet they remain insufficiently explored in existing machine learning studies [10], [11].

Another challenge in PM2.5 classification is the need for reliable multiclass prediction rather than simple regression. In practical environmental management, classifying air quality into interpretable categories such as good, moderate, and unhealthy is often more actionable than predicting concentration values alone. However, multiclass classification becomes difficult when class boundaries are close, particularly for intermediate categories where concentration overlap frequently occurs [5], [6]. Therefore, optimized learning algorithms are required to improve discrimination performance across all categories.

Hyperparameter tuning is essential for maximizing model generalization. Without proper optimization, models may suffer from underfitting, overfitting, or unstable performance on unseen samples. Grid Search is transparent for compact parameter grids, whereas randomized search can explore broader interacting parameter spaces more efficiently. Recent air-quality machine-learning studies also emphasize systematic model selection and optimization [12], [13], while optimization coupled with machine learning has been applied to other built-environment performance problems [14].

Based on these gaps, this study investigates PM2.5 multiclass classification using SVM, RF, and CatBoost models on a large dataset collected from wooden houses in highland and lowland areas of Indonesia. The dataset contains approximately 430,000 hourly observations of pollutant and meteorological variables, representing a realistic environmental monitoring scenario with heterogeneous conditions.

The contribution of this study is not the invention of a new learning algorithm, but the integration of four elements that are rarely evaluated together: i) large-scale sensor observations from wooden houses in contrasting highland and lowland contexts, ii) multiclass PM2.5 classification, iii) model-specific hyperparameter optimization, and iv) post-hoc SHAP interpretation of the optimized CatBoost model. Compared with recent PM2.5 and air-quality classification studies, this design explicitly emphasizes the residential wooden-house setting, contrasting measurement locations, model-specific optimization, and post-hoc explainability. Unlike studies centered on urban stations, regional forecasting, or non-residential datasets, the present work explicitly evaluates residential altitude context while retaining location as a categorical feature. Because each altitude is represented by a specific site, the study treats location as contextual information and does not interpret it as an isolated causal effect of altitude.

Recent related studies have addressed explainable air-quality classification using granular rule extraction [2], imbalanced multiclass PM2.5 classification with boosting algorithms in an urban Indonesian setting [5], multiclass PM2.5 prediction using an improved adaptive boosting rotation forest [6], transformer-based air-quality classification [15], and regional reclassification and PM2.5 prediction in a subtropical environment [16]. However, these studies were conducted primarily at urban, regional, or non-residential scales and did not jointly examine wooden-house environments across contrasting altitudes while combining model-specific hyperparameter optimization with explainable artificial intelligence (AI). The present study addresses this gap by integrating highland and lowland wooden houses, approximately 430,000 observations, SVM, RF, and CatBoost classification, systematic optimization, and SHAP-based interpretation within a single residential monitoring framework.

Therefore, this study aims to evaluate the comparative performance of SVM, RF, and CatBoost for multiclass PM2.5 classification in wooden houses, quantify the effect of model-specific hyperparameter optimization, interpret the optimized CatBoost model using SHAP, and assess the extent to which residential location provides contextual information for intelligent air-quality monitoring.

2. METHOD

This study employed a quantitative experimental approach to classify PM2.5 concentrations in wooden houses located in highland and lowland areas of Indonesia using machine-learning techniques. As illustrated in Figure 1, the workflow comprised environmental data collection, data cleaning and imputation, PM2.5 class labeling, an 80:20 stratified train-test split, model training, model-specific hyperparameter optimization with five-fold cross-validation, independent test evaluation, and best-model comparison conditions [6], [15], [17]-[19]. Explainable AI analysis was then conducted for the optimized CatBoost model to investigate the contribution of environmental and contextual features [20]-[22].

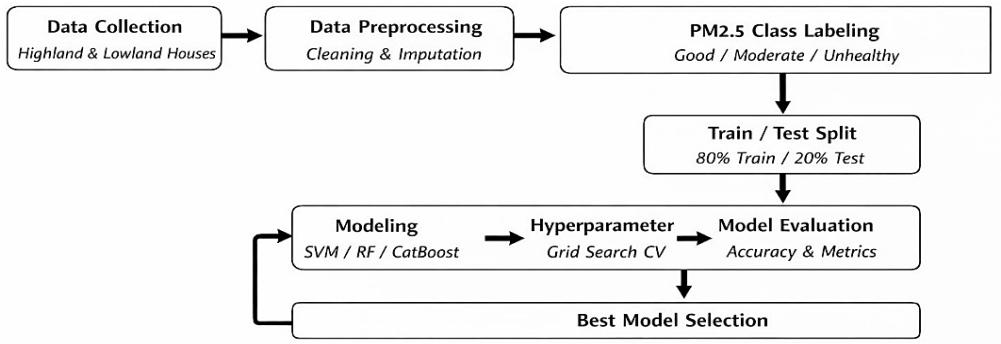


Figure 1. Research workflow

2.1. Dataset

The dataset contained approximately 430,000 hourly observations collected from wooden houses in Dieng (highland) and Rembang (lowland), Central Java, Indonesia. These locations were selected because they represent contrasting environmental and topographic conditions that may influence pollutant accumulation and dispersion. Each record consisted of PM2.5, PM10, NOx, globe temperature, digital humidity and temperature (DHT) humidity, AHT21 temperature, AHT21 relative humidity, eCO2, total volatile organic compounds (TVOC), wind speed, timestamp, and location variables. Similar multivariate air quality datasets have been effectively used in previous machine learning research [1], [23].

PM2.5 concentration values were transformed into three categorical classes: good (0–28.5 µg/m³), moderate (28.5–40.5 µg/m³), and unhealthy (>40.5 µg/m³). Multiclass categorization is more practical than regression outputs for environmental decision support systems and public health communication [5], [6].

2.2. Data preprocessing

The raw dataset was cleaned by standardizing column names, removing invalid characters from numerical attributes, and converting sensor readings into valid numeric format. The timestamp variable was transformed by extracting only the hourly component to capture diurnal pollution behavior, while date information was excluded. Records with missing hour or PM2.5 values were removed. Remaining missing numerical values were imputed using median values, while missing categorical labels were replaced with "Unknown." This preprocessing strategy was chosen because median imputation is robust against outliers and commonly used in environmental datasets [11], [13].

2.3. Classification models and optimization

Three supervised learning algorithms were evaluated: SVM, RF, and CatBoost. These models were selected because they represent different classification paradigms and have shown strong performance in environmental prediction tasks [5], [6], [20], [21], [24].

Hyperparameter optimization was model-specific. SVM and RF were tuned by grid search with five-fold stratified cross-validation. For CatBoost, the final experiment used CatBoost randomized_search with 25 candidate configurations over depth {4, 6, 8, 10}, learning_rate {0.03, 0.05, 0.07, 0.10}, l2_leaf_reg {1, 3, 5, 7, 10, 15}, bagging_temperature {0, 0.25, 0.5, 1.0}, random_strength {1, 2, 5, 10}, border_count {128, 254}, and iterations {300, 500, 700, 900}. CatBoost used multi class loss, total F1 evaluation, balanced class weights, and random seed 42. The best CatBoost configuration was depth=10, learning_rate=0.07, iterations=900, l2_leaf_reg=3, bagging_temperature=0.5, border_count=128, and random_strength=2.

The dataset was divided into training and testing sets using an 80:20 stratified split with random_state=42, resulting in 344,028 training observations and 86,008 independent test observations for the

reproduced CatBoost experiment. Hyperparameter selection and class-weight calculation used only the training partition, and the test set was reserved for final evaluation. Five-fold stratified cross-validation was applied during model selection to preserve class distribution and reduce selection bias [16], [25]-[27].

2.4. Performance evaluation

Model performance was evaluated using accuracy, precision, recall, f1-score, matthews correlation coefficient (MCC), and confusion matrix. Since the dataset contained multiple classes with different distributions, macro F1-score and MCC were emphasized to provide balanced assessment across all categories [3], [15], [28], [29].

2.5. Reproducibility and SHAP-based explainability

The implementation used Python with NumPy, pandas, scikit-learn, CatBoost (>=1.2.7), Matplotlib, and SHAP. The implementation pseudocode was: i) clean and impute the sensor dataset; ii) create the stratified 80:20 split with random_state=42; iii) calculate class weights from the training partition; iv) perform five-fold model-specific hyperparameter search; v) refit the selected configuration on the full training partition; vi) evaluate once on the independent test set; and vii) calculate post-hoc explanations. Figure 1 provides the corresponding implementation flowchart.

SHAP was applied to the optimized CatBoost model because CatBoost showed the largest improvement after optimization and natively supports categorical location information and SHAP values. Native CatBoost ShapValues were computed from 5,000 observations randomly sampled from the independent test set using random_state=42. Global feature importance was calculated as the mean absolute SHAP value across observations and classes, while class-specific mean absolute SHAP values were used to examine differences among good, moderate, and unhealthy predictions. SHAP values are interpreted as explanations of model predictions rather than evidence of causal environmental effects [2], [10], [30].

3. RESULTS AND DISCUSSION

To place the experimental results in context, Table 1 compares the present study with representative recent PM2.5 and air-quality classification research and highlights the specific gap addressed by the altitude-specific wooden-house setting, model optimization, and SHAP-based interpretation.

Table 1. Positioning of the present study relative to recent PM2.5 and air-quality classification research

Study	Application context	Method	Primary emphasis	Gap addressed by the present study
[2]	Air-quality classification	Explainable granular rule extraction.	Interpretability	No altitude-specific wooden-house setting.
[5]	Jakarta PM2.5 concentration levels	AdaBoost, XGBoost, CatBoost, and LightGBM.	Imbalanced multiclass PM2.5 classification	Urban context; no wooden-house altitude contrast.
[6]	PM2.5 concentration categories	Improved adaptive boosting rotation forest.	Multiclass prediction	No residential altitude-specific environment.
[15]	Air-quality classification	Double-output vision transformer.	Transformer-based classification	Different data representation; no wooden-house altitude context.
[16]	Subtropical air-quality zones	Machine-learning reclassification and PM2.5 prediction.	Spatial reclassification and prediction	Regional rather than household-level context.
Present study	Highland and lowland wooden houses; ~430,000 observations	SVM, RF, CatBoost+model-specific optimization+SHAP.	Residential multiclass classification with interpretable site/altitude context	Integrates housing context, altitude contrast, optimization, and explainable artificial intelligence (XAI).

3.1. Overall performance before and after optimization

Table 2 summarizes the classification results before and after hyperparameter optimization. Model-specific search improved all three classifiers, indicating that parameter selection materially affected predictive performance and generalization.

RF achieved the best overall performance with 98.19% accuracy, a 0.9763 macro F1-score, and 0.9655 MCC. Relative to its already strong baseline, optimization increased accuracy by 0.24 percentage points, macro F1 by 0.0040, and MCC by 0.0039. The modest gain reflects limited remaining headroom at baseline, while tuning tree number, depth, feature subsampling, and split/leaf constraints improved the bias-variance balance and decorrelation among trees. This behavior is consistent with the robustness of ensemble bagging on noisy, nonlinear environmental data [12], [21], [26], [31], [32].

Table 2. Classification performance before and after hyperparameter optimization

Model	Stage	Best parameters	Accuracy (%)	Macro F1-score	MCC
SVM radial basis function (RBF)	Before	Default setting	94.41	0.9290	0.8993
SVM (RBF)	After	C=0.3149; gamma=0.6351	94.71	0.9339	0.9049
RF	Before	n_estimators=300	97.95	0.9723	0.9616
RF	After	n_estimators ≈600; max_depth=28; max_features=sqrt; min_samples_split=4; min_samples_leaf=1	98.19	0.9763	0.9655
CatBoost	Before	depth=6; learning_rate=0.08; iterations=500	95.84	0.9473	0.9244
CatBoost	After	depth=10; learning_rate=0.07; iterations=900; l2_leaf_reg=3; bagging_temperature=0.5; border_count=128; random_strength=2	97.12	0.9630	0.9471

CatBoost showed the largest optimization gain: accuracy increased from 95.84% to 97.12% (+1.28 percentage points), macro F1 from 0.9473 to 0.9630, and MCC from 0.9244 to 0.9471. The selected combination of deeper trees, a moderate learning rate, 900 boosting iterations, L2 regularization, bagging temperature, border count, and random strength allowed the boosting process to refine difficult boundary regions while controlling model complexity [6], [22].

SVM improved from 94.41% to 94.71% accuracy and from 0.8993 to 0.9049 MCC. The smaller gain suggests that hyperparameter tuning alone could not fully match the flexibility of the tree ensembles on this large heterogeneous dataset. Although the RBF kernel provides nonlinear decision functions, SVM training and decision behavior can be more sensitive to kernel parameters and data scale than ensemble-tree methods in large air-quality applications [20], [33].

Overall, these findings indicate that ensemble-based methods were more suitable than SVM for PM2.5 classification in residential environments. RF provided the most stable performance, while CatBoost benefited the most from optimization.

3.2. Class-wise performance after optimization

After optimization, all models achieved strong performance for the good and unhealthy classes, whereas the moderate class remained the most challenging category. This pattern indicates that samples located at the lower and higher PM2.5 ranges were easier to separate because their pollutant and meteorological characteristics were more distinct, while intermediate observations showed greater overlap among classes. RF delivered the most balanced class-wise performance, with F1-scores of 0.9904 for good, 0.9597 for moderate, and 0.9788 for unhealthy classes, corresponding to 98.19% accuracy and 0.9655 MCC. Its superior results are mainly attributed to the ensemble bagging mechanism, where multiple decision trees capture nonlinear interactions from different feature subsets and combine them through majority voting. This structure increases robustness to noise and improves discrimination in overlapping environmental data patterns [21].

CatBoost also showed competitive performance, with F1-scores of 0.9841, 0.9332, and 0.9717 for the good, moderate, and unhealthy classes, respectively, together with 97.12% accuracy and 0.9471 MCC. Gradient boosting sequentially fits trees to the loss gradients, allowing later trees to refine residual errors in difficult regions of feature space. Nevertheless, the moderate-class F1-score remained below that of RF, indicating that transitional observations near adjacent class boundaries were still harder to discriminate [5], [24], [34].

SVM produced lower results, particularly for the moderate class, with F1-scores of 0.9694, 0.8841, and 0.9483, together with 94.71% accuracy and 0.9049 MCC. This indicates that the RBF decision function retained more overlap among intermediate samples than the ensemble models, even after tuning [20], [33].

The lower performance of all models in the moderate category is expected because its threshold range (28.5–40.5 µg/m³) lies between adjacent classes, causing overlapping feature patterns and less distinct decision boundaries [6], [32]. Overall, RF provided the most consistent class-level classification, while CatBoost showed strong capability in identifying unhealthy air quality conditions [5], [35]-[37].

3.3. Confusion matrix insights

Figure 2 presents the confusion matrices of SVM, RF, and CatBoost, where Figure 2(a) before hyperparameter optimization, and Figure 2(b) presents the corresponding matrices after optimization. In both subfigures, diagonal entries represent correctly classified observations, whereas off-diagonal entries indicate misclassifications between PM2.5 classes.

The confusion matrices show that the good and unhealthy classes exhibit stronger diagonal concentrations, whereas the moderate class produces proportionally more off-diagonal errors. This pattern is consistent with the moderate concentration range (28.5–40.5 µg/m³), which lies between the adjacent classes and is associated with overlapping feature patterns.

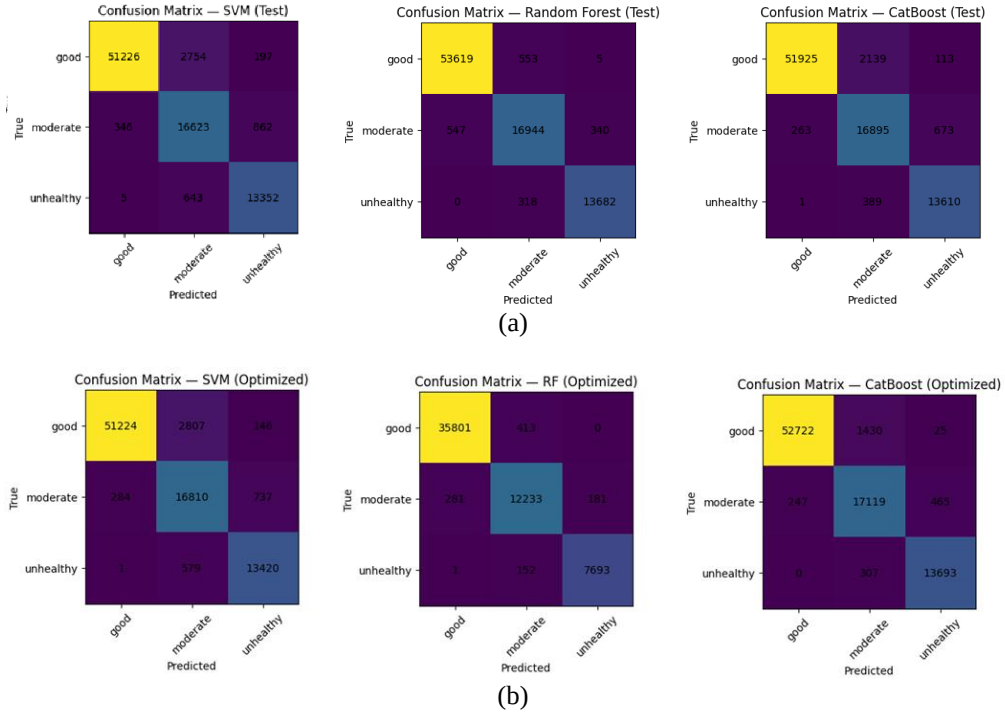


Figure 2. Confusion matrices of classification results; (a) SVM, RF, and CatBoost before optimization and (b) SVM, RF, and CatBoost after optimization

For CatBoost, the total number of test-set errors decreased from 3,578 before optimization to 2,474 after optimization, a 30.9% reduction. Good-to-moderate errors decreased from 2,139 to 1,430 (33.1%), moderate-to-unhealthy errors from 673 to 465 (30.9%), and unhealthy-to-moderate errors from 389 to 307 (21.1%). Moderate-class recall consequently increased from 94.75% to 96.01%. These changes show that the optimized CatBoost parameters improved separation particularly around adjacent class boundaries rather than only increasing the aggregate accuracy.

RF produced the strongest overall confusion-matrix pattern and the highest aggregate metrics after optimization. Its tuned depth, feature subsampling, and split constraints reduce variance while preserving diverse decision paths across trees, which is advantageous when environmental features interact nonlinearly. Because the RF baseline was already strong, its numerical gain after tuning was smaller than CatBoost's. SVM retained comparatively more confusion around the moderate class, consistent with its lower macro F1-score [26], [32], [37]. These results confirm that RF was not only the best model in overall metrics, but also the most reliable classifier for minimizing practical classification errors in residential PM2.5 monitoring.

3.4. Shapley additive explanations-based explainable artificial intelligence analysis

To improve model interpretability, SHAP analysis was performed on the optimized CatBoost classifier using 5,000 observations from the independent test set. Figure 3 presents the global mean absolute SHAP values. PM10 was by far the dominant predictor (2.9566), more than ten times the contribution of the next-ranked feature, eCO2 (0.2914), followed by TVOC (0.2785), NOx (0.1905), DHT humidity (0.1711), hour (0.1706), AHT21 relative humidity (0.1699), and location (0.1696). This dominance indicates that PM10 provides the strongest information for the model's PM2.5 class decisions, but the SHAP result should not be interpreted as a causal effect because PM10 and PM2.5 are physically related particulate fractions [2], [10].

Figure 3 global SHAP feature importance of the optimized CatBoost model for multiclass PM2.5 classification. Class-specific SHAP importance further clarified the moderate-class difficulty. PM10 mean |SHAP| was 3.8378 for the good class and 3.2245 for the unhealthy class, but only 1.8073 for the moderate class. In the moderate category, humidity, NOx, and hour therefore contributed proportionally more to the decision than in the two extreme classes, consistent with the lower moderate F1-score. Location contributed 0.1942, 0.1277, and 0.1870 for the good, moderate, and unhealthy classes, respectively. These values show that site/altitude context provides secondary information, but because the two altitude levels correspond to specific locations, SHAP cannot separate altitude from other site-specific characteristics.

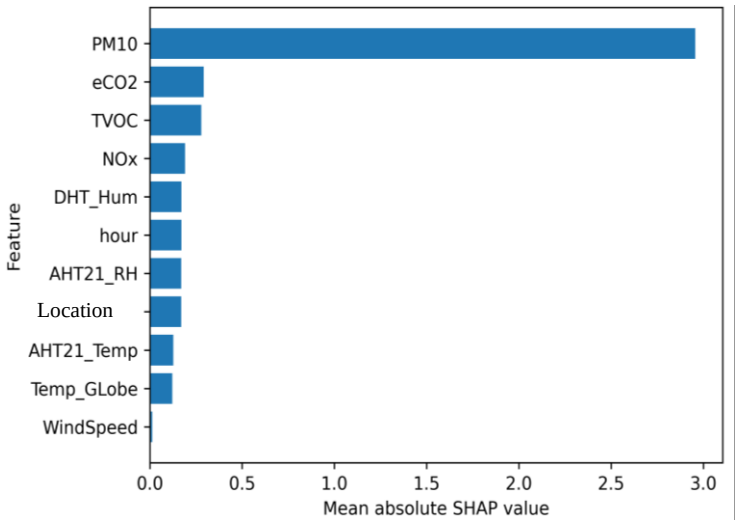


Figure 3. Global SHAP feature importance of the optimized CatBoost model for multiclass PM2.5 classification

3.5. Impact of optimization and practical implications

The optimization process had a clear positive impact on predictive accuracy, macro F1-score, and MCC, but the magnitude of improvement depended on model structure and baseline strength. The results therefore support model-specific optimization rather than a single tuning strategy for all algorithms. This distinction is important for reproducibility because SVM/RF grid search and CatBoost randomized search explore parameter spaces differently [4], [13], [27].

Recent studies have explored ensemble learning, transformer-based classification, and explainable AI for environmental data [2], [15], [27], [28]. The present results do not claim algorithmic novelty over these architectures; instead, they show that conventional tree ensembles remain highly competitive for large tabular sensor datasets. A direct transformer benchmark was not performed, so competitiveness against transformer architectures should be regarded as a target for future experimental comparison rather than an established result.

These findings directly address the importance of optimization in machine learning-based environmental monitoring. Optimization should not be viewed as an optional step, but as a necessary procedure to obtain fair model comparison and maximum predictive reliability [1], [4], [13], [14], [27].

From a practical perspective, the optimized RF is the strongest candidate for subsequent edge/IoT validation because it achieved the highest predictive performance. However, deployment readiness cannot yet be claimed because inference latency, model size, memory use, processor load, and energy consumption were not measured on embedded hardware. Future deployment studies should therefore benchmark these quantities alongside accuracy so that predictive gains can be evaluated against computational and energy costs [5], [14], [27], [36], [37].

Overall, the findings show that ensemble learning is effective for this large heterogeneous residential dataset, while SHAP adds a transparent explanation layer to the optimized CatBoost model. RF provided the best predictive balance, CatBoost showed the largest response to optimization, and location supplied secondary contextual information. These results extend PM2.5 classification beyond conventional urban-station settings while also defining clear limits on claims about altitude and real-time deployment.

3.6. Limitations, robustness, and scalability

Several limitations should be considered when interpreting the results. First, the dataset represents two residential sites, one highland and one lowland, so the categorical location variable combines altitude with other site-specific characteristics such as local emissions, microclimate, and household context [10]. Consequently, the contribution of location in SHAP should not be interpreted as an isolated causal altitude effect. Second, the random stratified split can place temporally adjacent observations in both training and test sets. This limitation is particularly relevant for sequential environmental sensor data because temporal autocorrelation may produce optimistic performance estimates when observations from similar periods occur in both subsets [38], [39]. Therefore, the reported metrics characterize within-distribution discrimination more directly than strict future-time or unseen-site transfer.

Robustness to systematic sensor calibration error, structured missingness, and environmental disturbances was not stress-tested experimentally. Low-cost air-quality sensors may exhibit measurement bias, calibration instability, and performance variation under different environmental conditions, indicating the importance of appropriate calibration and validation procedures [38], [39]. Median imputation addresses missing values during preprocessing but does not establish robustness to sensor drift, sustained data loss, or changing operating conditions [39], [40]. Future validation should introduce controlled perturbations such as ±5% and ±10% sensor bias, 5–20% missing-data scenarios, noise disturbances, blocked temporal validation, and leave-location-out testing. Multi-season, multi-climate, and multi-altitude datasets with several sites per altitude band are also required to evaluate scalability and external validity [38].

Computational efficiency and energy consumption were also outside the present experimental scope. These aspects are important for practical IoT and edge-computing deployment because real-time environmental monitoring requires a balance among prediction accuracy, latency, computational resources, scalability, and energy consumption [41]. Future work should report training time, inference latency, model size, RAM/CPU utilization, and energy consumption per fixed number of predictions on representative edge hardware. The best ensemble models should also be compared directly with recent transformer-based air-quality classification approaches [15] and explainable artificial intelligence architectures [30]. These additions would allow predictive accuracy, interpretability, robustness, and deployment cost to be assessed within a single framework.

4. CONCLUSION

This study evaluated SVM, RF, and CatBoost for multiclass PM2.5 classification in wooden houses located in contrasting highland and lowland contexts using approximately 430,000 observations. Model-specific hyperparameter optimization improved all classifiers. RF achieved the best overall performance with 98.19% accuracy, a 0.9763 macro F1-score, and 0.9655 MCC, followed by CatBoost and SVM.

Class-level analysis confirmed that the moderate category was the most difficult because it lies between adjacent concentration classes and exhibits greater feature overlap. In CatBoost, optimization reduced total test errors by 30.9%, including substantial reductions in good-to-moderate and moderate-to-unhealthy transitions, demonstrating improved boundary discrimination rather than only a higher aggregate score.

SHAP analysis added interpretability to the optimized CatBoost model. PM10 was the dominant predictor (global mean |SHAP|=2.9566), followed by eCO2 and TVOC, while humidity, temporal information, and location provided secondary contributions. The weaker PM10 dominance in the moderate class helps explain its lower F1-score. Location contributed contextual information, but the study does not interpret this contribution as a causal altitude effect because altitude and site are not independently replicated.

Future work should prioritize multi-site and multi-season data collection, temporally blocked and leave-location-out validation, controlled sensor-error and missing-data robustness experiments, edge-device latency and energy benchmarks, and direct comparison with recent transformer-based and explainable ensemble approaches. These extensions are required before claims of broad geographic generalization or real-time edge deployment can be established.

ACKNOWLEDGEMENTS

The authors acknowledge the Directorate of Research, Technology, and Community Service (DRTPM), LLDIKTI Region VI, and LP3M Universitas Sains Al-Qur'an for institutional and research support provided under the research grant/contract listed below.

FUNDING INFORMATION

This research was funded by DRTPM, Ministry of Higher Education, Science, and Technology, Republic of Indonesia with contract number 127/C3/DT.05.00/PL/2025, 027/LL6/PL/AL.04/2025 (LLDIKTI 6), A.2.02/LP3M-UNSIQ/PFR/2025 (LP3M UNSIQ).

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Muhamad Fuat Asnawi	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓	✓
Muhammad Alif Muwafiq Baihaqy	✓		✓	✓			✓			✓	✓		✓	✓
Adinda Septi Hendriani		✓				✓		✓	✓		✓		✓	✓
Nulngafan			✓		✓					✓	✓			
Hermawan		✓				✓		✓	✓	✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Kaggle at https://doi.org/10.34740/kaggle/dsv/13204736. The dataset has been deposited by the author and can be freely downloaded for research and academic purposes.

REFERENCES

[1] A. Masood *et al.*, "Improving PM2.5 prediction in New Delhi using a hybrid extreme learning machine coupled with snake optimization algorithm," *Scientific Reports*, vol. 13, no. 1, p. 21057, Nov. 2023, doi: 10.1038/s41598-023-47492-z.

[2] I. Jairi, S. Ben-Othman, L. Canivet, and H. Zgaya-Biau, "Explainable based approach for the air quality classification on the granular computing rule extraction technique," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108096, Jul. 2024, doi: 10.1016/j.engappai.2024.108096.

[3] M. Kaya and İ. Ö. Bucak, "Time series analysis and prediction of PM2.5 pollution concentration in Igdir province with deep learning models," *Atmospheric Environment*, vol. 361, p. 121491, Nov. 2025, doi: 10.1016/j.atmosenv.2025.121491.

[4] T. Peng *et al.*, "Research and application of a novel selective stacking ensemble model based on error compensation and parameter optimization for AQI prediction," *Environmental Research*, vol. 247, p. 118176, Apr. 2024, doi: 10.1016/j.envres.2024.118176.

[5] T. Toharudin *et al.*, "Boosting Algorithm to Handle Unbalanced Classification of PM2.5 Concentration Levels by Observing Meteorological Parameters in Jakarta-Indonesia Using AdaBoost, XGBoost, CatBoost, and LightGBM," *IEEE Access*, vol. 11, pp. 35680–35696, 2023, doi: 10.1109/ACCESS.2023.3265019.

[6] T. Deng *et al.*, "Multi-classification prediction of PM2.5 concentration based on improved adaptive boosting rotation forest," *Journal of Environmental Chemical Engineering*, vol. 12, no. 6, p. 114658, Dec. 2024, doi: 10.1016/j.jece.2024.114658.

[7] Z. Zhang, Y. Zeng, and K. Yan, "A hybrid deep learning technology for PM2.5 air quality forecasting," *Environmental Science and Pollution Research*, vol. 28, no. 29, pp. 39409–39422, Aug. 2021, doi: 10.1007/s11356-021-12657-8.

[8] P. P. Vignesh, J. H. Jiang, and P. Kishore, "Predicting PM2.5 Concentrations Across USA Using Machine Learning," *Earth and Space Science*, vol. 10, no. 10, Oct. 2023, doi: 10.1029/2023EA002911.

[9] J. Dutta and S. Roy, "OccupancySense: Context-based indoor occupancy detection & prediction using CatBoost model," *Applied Soft Computing*, vol. 119, p. 108536, Apr. 2022, doi: 10.1016/j.asoc.2022.108536.

[10] S. Wang *et al.*, "An interpretable spatially weighted machine learning approach for revealing spatial nonstationarity impacts of the built environment on air pollution," *Building and Environment*, vol. 280, p. 113150, Jul. 2025, doi: 10.1016/j.buildenv.2025.113150.

[11] M. Rahaman *et al.*, "Combining deep learning and machine learning techniques to track air pollution in relation to vegetation cover utilizing remotely sensed data," *Journal of Environmental Management*, vol. 376, p. 124323, Mar. 2025, doi: 10.1016/j.jenvman.2025.124323.

[12] M. Méndez, M. G. Merayo, and M. Núñez, "Machine learning algorithms to forecast air quality: a survey," *Artificial Intelligence Review*, vol. 56, no. 9, pp. 10031–10066, Sep. 2023, doi: 10.1007/s10462-023-10424-4.

[13] F. Yang, X. Jiang, and Z. Chen, "Air Quality Index Prediction Model Based on Multiple Attention Mechanisms and Hyperparameter Optimization," in *2023 3rd International Conference on Electronic Information Engineering and Computer Science, EIECS 2023*, Sep. 2023, pp. 603–606, doi: 10.1109/EIECS59936.2023.10435483.

[14] Z. Chen, Y. Cui, H. Zheng, and Q. Ning, "Optimization and prediction of energy consumption, light and thermal comfort in teaching building atriums using NSGA-II and machine learning," *Journal of Building Engineering*, vol. 86, p. 108687, Jun. 2024, doi: 10.1016/j.jobe.2024.108687.

[15] Z. Wang, Y. Yang, and S. Yue, "Air Quality Classification and Measurement Based on Double Output Vision Transformer," *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 20975–20984, Nov. 2022, doi: 10.1109/JIOT.2022.3176126.

[16] N. Akhtar and J. Y. Han, "Machine learning –driven reclassification of air quality zones and PM2.5 prediction: A case study in a subtropical region," *Urban Climate*, vol. 65, p. 102766, Feb. 2026, doi: 10.1016/j.uclim.2025.102766.
[17] S. A. Syed, N. Aditi, S. Prajwal, A. Kodipalli, T. Rao, and B. R. Rohini, "Prediction and Classification of Contamination of Pollutants in Air Using Machine Learning Algorithms," in *Proceedings - 2023 International Conference on Computational Intelligence for Information, Security and Communication Applications, CIISCA 2023*, Jun. 2023, pp. 236–240, doi: 10.1109/CIISCA59740.2023.00053.
[18] R. S. Rao, L. R. Kalabarige, M. R. Holla, and A. K. Sahu, "Multimodal Imputation-Based Multimodal Autoencoder Framework for AQI Classification and Prediction of Indian Cities," *IEEE Access*, vol. 12, pp. 108350–108363, 2024, doi: 10.1109/ACCESS.2024.3438573.
[19] W. Choi, C. H. Ho, and Y. Lee, "Temporal pattern classification of PM2.5 chemical compositions in Seoul, Korea using K-means clustering analysis," *Science of the Total Environment*, vol. 927, p. 172157, Jun. 2024, doi: 10.1016/j.scitotenv.2024.172157.
[20] S. Ketu and P. K. Mishra, "Scalable kernel-based SVM classification algorithm on imbalance air quality data for proficient healthcare," *Complex and Intelligent Systems*, vol. 7, no. 5, pp. 2597–2615, Oct. 2021, doi: 10.1007/s40747-021-00435-5.
[21] M. Asgari, W. Yang, and M. Farnaghi, "Spatiotemporal data partitioning for distributed random forest algorithm: Air quality prediction using imbalanced big spatiotemporal data on spark distributed framework," *Environmental Technology and Innovation*, vol. 27, p. 102776, Aug. 2022, doi: 10.1016/j.eti.2022.102776.
[22] Y. Ding, Z. Chen, W. Lu, and X. Wang, "A CatBoost approach with wavelet decomposition to improve satellite-derived high-resolution PM2.5 estimates in Beijing-Tianjin-Hebei," *Atmospheric Environment*, vol. 249, p. 118212, Mar. 2021, doi: 10.1016/j.atmosenv.2021.118212.
[23] M. Su, H. Liu, C. Yu, and Z. Duan, "A novel AQI forecasting method based on fusing temporal correlation forecasting with spatial correlation forecasting," *Atmospheric Pollution Research*, vol. 14, no. 4, p. 101717, Apr. 2023, doi: 10.1016/j.apr.2023.101717.
[24] Y. Su, "Prediction of air quality based on Gradient Boosting Machine Method," in *Proceedings - 2020 International Conference on Big Data and Informatization Education, ICBDIE 2020*, Apr. 2020, pp. 395–397, doi: 10.1109/ICBDIE50010.2020.00099.
[25] S. Chen, X. Xi, M. Zhong, R. Yang, and M. E. Orchard, "Handling mislabeled data in fault diagnosis: A graph-assisted random forest approach," *Neurocomputing*, vol. 671, p. 132669, Mar. 2026, doi: 10.1016/j.neucom.2026.132669.
[26] F. Alzu'bi, A. Al-Rawabdeh, and A. Almagbile, "Predicting air quality using random forest: A case study in Amman-Zarqa," *Egyptian Journal of Remote Sensing and Space Science*, vol. 27, no. 3, pp. 604–613, Sep. 2024, doi: 10.1016/j.ejrs.2024.07.004.
[27] M. Mubeen *et al.*, "Smart prediction and optimization of air quality index with artificial intelligence," *Journal of Environmental Sciences (China)*, vol. 158, pp. 761–775, Dec. 2025, doi: 10.1016/j.jes.2025.02.041.
[28] L. Dang *et al.*, "A novel classifier ensemble for recognition of multiple indoor air contaminants by an electronic nose," *Sensors and Actuators, A: Physical*, vol. 207, pp. 67–74, Mar. 2014, doi: 10.1016/j.sna.2013.12.029.
[29] L. Roy, A. Banerjee, R. Ghosh, K. Bhattacharya, and S. Pal, "Combined Experimental and Simulation Studies for Potential Classification of Inorganic and Organic Suspended Particulates in Indoor Air," *IEEE Sensors Letters*, vol. 8, no. 8, pp. 1–4, Aug. 2024, doi: 10.1109/LSENS.2024.3427366.
[30] M. S. S. Sunder, V. A. Tikkiwal, A. Kumar, and B. Tyagi, "Unveiling the Transparency of Prediction Models for Spatial PM2.5 over Singapore: Comparison of Different Machine Learning Approaches with eXplainable Artificial Intelligence," *AI (Switzerland)*, vol. 4, no. 4, pp. 787–811, Sep. 2023, doi: 10.3390/ai4040040.
[31] W. Hernandez, A. Mendez, R. Zalakeviciute, and A. M. Diaz-Marquez, "Analysis of the Information Obtained from PM2.5 Concentration Measurements in an Urban Park," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 9, pp. 6296–6311, Sep. 2020, doi: 10.1109/TIM.2020.2966360.
[32] R. E. A. Durán, M. Cortez-Huerta, R. S. Echeverría, G. F. García, E. C. Valdez, and J. D. W. Kahl, "Random Forest approach to air quality assessment for port emissions in a coastal area in the Gulf of Mexico," *Marine Pollution Bulletin*, vol. 222, p. 118765, Jan. 2026, doi: 10.1016/j.marpolbul.2025.118765.
[33] H. Balogun, H. Alaka, and C. N. Egwim, "Boruta-grid-search least square support vector machine for NO2 pollution prediction using big data analytics and IoT emission sensors," *Applied Computing and Informatics*, vol. 21, no. 1–2, pp. 101–113, Jan. 2025, doi: 10.1108/ACI-04-2021-0092.
[34] T. Zhang, W. He, H. Zheng, Y. Cui, H. Song, and S. Fu, "Satellite-based ground PM2.5 estimation using a gradient boosting decision tree," *Chemosphere*, vol. 268, p. 128801, Apr. 2021, doi: 10.1016/j.chemosphere.2020.128801.
[35] T. D. Morapedi and I. C. Obagbuwa, "Air pollution particulate matter (PM2.5) prediction in South African cities using machine learning techniques," *Frontiers in Artificial Intelligence*, vol. 6, Oct. 2023, doi: 10.3389/frai.2023.1230087.
[36] C. N. Egwim *et al.*, "Ensemble of ensembles for fine particulate matter pollution prediction using big data analytics and IoT emission sensors," *Journal of Engineering, Design and Technology*, vol. 23, no. 2, pp. 640–665, Feb. 2025, doi: 10.1108/JEDT-07-2022-0379.
[37] A. Gururaj, V. Nagaraj, A. S. Rajesh, K. N. Achyuth, M. U. Somesh, and B. A. Dubay, "Air Quality Prediction and Analysis using Machine Learning," in *Proceedings - 2024 5th International Conference on Mobile Computing and Sustainable Informatics, ICMCSI 2024*, Jan. 2024, pp. 543–550, doi: 10.1109/ICMCSI61536.2024.00086.
[38] S. Sousan, R. Wu, C. Popoviciu, S. Fresquez, and Y. M. Park, "Advancing low-cost air quality monitor calibration with machine learning methods," *Environmental Pollution*, vol. 374, p. 126191, Jun. 2025, doi: 10.1016/j.envpol.2025.126191.
[39] J. Dong, N. Goodman, A. Carre, and P. Rajagopalan, "Calibration and validation-based assessment of low-cost air quality sensors," *Science of the Total Environment*, vol. 977, p. 179364, May 2025, doi: 10.1016/j.scitotenv.2025.179364.
[40] C. Jin, Y. Wang, T. Li, and Q. Yuan, "Global validation and hybrid calibration of CAMS and MERRA-2 PM2.5 reanalysis products based on OpenAQ platform," *Atmospheric Environment*, vol. 274, 2022, doi: 10.1016/j.atmosenv.2022.118972.
[41] E. Alabdulkreem *et al.*, "Innovative IoT and blockchain integration for real-time urban air quality monitoring and autonomous response system," *Sustainable Computing: Informatics and Systems*, vol. 48, p. 101250, Dec. 2025, doi: 10.1016/j.suscom.2025.101250.

BIOGRAPHIES OF AUTHORS



Muhamad Fuat Asnawi is a researcher at Universitas Sains Al-Qur'an (UNSIQ), Wonosobo, Indonesia. He received his bachelor's degree in Information Technology (2013) and master's degree in Management Information Systems (2015) from Universitas Budi Luhur, Indonesia, and a master's degree in Informatics from Universitas AMIKOM Yogyakarta (2025). He is currently a doctoral student in Informatics at Universitas AMIKOM Yogyakarta. His research interests include data science, machine learning, air quality monitoring, intelligent systems, and information systems management. He has published several works on environmental data analysis and intelligent computing. He can be contacted at email: fuatasnawi@unsiq.ac.id.



Muhammad Alif Muwafiq Baihaqy is a researcher and faculty member whose academic background is in Informatics. He earned his bachelor's degree in Informatics from Universitas Sains Al-Qur'an in 2009, and later obtained his master's degree in Informatics from STMIK AMIKOM in 2016. His research interests include (based on his publications) areas such as software engineering, information systems, data analytics, and related fields. He can be contacted at email: aviq.baihaqy@gmail.com.



Adinda Septi Hendriani is a researcher at Universitas Sains Al-Qur'an (UNSIQ), Indonesia. She earned her bachelor's and master's degrees in Architecture from Universitas Diponegoro in 2009 and 2012, respectively. Her research interests include urban and settlement design, thermal comfort, vernacular architecture, and the study of traditional wooden houses in Indonesia. She has published works focusing on architectural heritage, sustainable housing, and the preservation of wooden residential buildings. She can be contacted at email: adinda@unsiq.ac.id.



Nulngafan is a researcher at Universitas Sains Al-Qur'an (UNSIQ), Indonesia. He received his bachelor's degree in Computer Science from UNSIQ, Wonosobo, in 2008 and his master's degree in Information Systems (CIO specialization) from Universitas AMIKOM Yogyakarta in 2013. His research interests include enterprise information systems, IT governance, management of information technology, and digital transformation. He has published several scholarly articles in relevant fields. He can be contacted at email: affan@unsiq.ac.id.



Prof. Dr. Hermawan is a professor in the Department of Architecture, Universitas Sains Al-Qur'an, He received the bachelor's degree in Architecture Engineering in 1999 from Universitas Diponegoro (UNDIP) Semarang, Master of Management at UNDIP in 2002, master of Architecture Engineering at UNDIP in 2007. Doctor of Architecture at UNDIP in 2020, Doctor of Architecture at Universitas Parahyangan Bandung in 2022. He received the academic leader award from LLDIKTI Regional 6 in 2024. Several competitive researches from DIKTI were obtained in 2015-2016 with the title Building Adaptive Thermal Comfort Theory in Extreme Tropical Areas, 2017-2019 with the title Comfortable and Healthy House Model in Mountainous Tropical Areas, 2022 with the title Indoor Air Quality (IAQ) Prediction Model for Wooden Houses in the Highlands and Lowlands to Create Energy-Efficient Buildings. Several articles in internationally reputable journals have also been published in the Journal of Applied Engineering Science (JAES), International Journal on Advanced Science, Engineering and Information Technology (IJASEIT), Journal of Applied Science and Engineering (JASE), energies, and sustainability. His research interests include thermal comfort architecture and local wisdom. He can be contacted at email: hermawan@unsiq.ac.id.