TIERS Information Technology Journal
Vol.7, No.1, June 2026, pp. 17~31
P-ISSN: 2723-4533 / E-ISSN: 2723-4541
DOI: https://doi.org/10.38043/tiers.v7i1.7558

Safety-Oriented Air Quality Index Classification for Imbalanced Data Using Optimized Boosting Models with Optuna and Oversampling

Made Yudi Dwipayana[1], Gede Angga Pradipta[2], Dandy Pramana Hostiadi[3]

232012016@stikom-bali.ac.id[1], angga_pradipta@stikom-bali.ac.id[2], dandy@stikom-bali.ac.id[3]
[1] Magister Program, Department of Magister Information Systems, Institut Teknologi dan Bisnis STIKOM Bali, Indonesia
[2,3] Department of Magister Information Systems, Institut Teknologi dan Bisnis STIKOM Bali, Indonesia

ABSTRACT

Air Quality Index (AQI) classification is essential for communicating environmental health risks. However, hazardous air conditions occur far less frequently than normal conditions, challenging conventional classification models. This study investigates multi-class AQI classification using the "Global Air Quality 2025" dataset, comprising 52,704 observations with an extreme class imbalance ratio of approximately 1:173. Under such conditions, conventional accuracy metrics often mask systemic failures in detecting critical minority classes. To address potential data leakage present in previous approaches, this research implements a rigorous cross-validation architecture combined with an independent 20% hold-out test set. The methodology employs an Ablation Study to systematically isolate the impacts of Optuna hyperparameter tuning guided by Macro F1-Score and oversampling techniques (SMOTE and ADASYN). The results demonstrate that the proposed Hybrid-SMOTE LightGBM configuration successfully balances hazard detection sensitivity with global stability. On the unseen hold-out set, the optimal model achieved a Macro F1-Score of 0.8079, an accuracy of 92.80%, and a ROC-AUC of 0.9847. Crucially, the model delivered a 65.12% recall for the critical Unhealthy minority class, a nearly 40% improvement over the baseline. Error profile analysis confirmed the model's safety-oriented robustness, as 97.6% of peak hazardous events were either accurately classified or safely constrained to the adjacent warning category, minimizing catastrophic misclassifications. These findings prove that reliable detection of environmental hazards requires safety-oriented per-class evaluation and strict validation frameworks, as reliance on aggregate global metrics leads to dangerously misleading performance assessments.

Keywords: Air Quality Index; Extreme Class Imbalance; Ensemble Boosting; SMOTE; Minority Class Detection; Robustness Evaluation

Article Info

Received : 21-01-2026
Revised : 18-03-2026
Accepted : 20-06-2026

Correspondence Author:

1. INTRODUCTION

Air pollution remains one of the most pressing environmental and public health challenges in urban regions worldwide [1]. Long-term exposure to elevated levels of particulate matter and gaseous pollutants has been consistently associated with respiratory diseases, cardiovascular complications, and increased mortality. As metropolitan areas continue to expand, reliable and timely air quality monitoring systems become critical components of public health infrastructure and environmental governance [2]. The Air Quality Index (AQI) provides a standardized mechanism for translating pollutant concentrations into categorical health risk levels, enabling rapid communication of risk severity to policymakers and the public [3]. With the advancement of sensor networks and data-driven monitoring platforms, machine learning techniques have been increasingly deployed to automate AQI classification and enhance early warning capabilities [4].

Despite these technological advances, AQI classification may become challenging due to class imbalance, where hazardous air quality conditions occur far less frequently than normal conditions, leading to skewed class distributions [5]. Such imbalance may reduce the effectiveness of machine learning models in identifying minority AQI categories, particularly when imbalance-handling strategies are insufficient. To address this issue, several studies have explored resampling and ensemble-based strategies. For instance [6] integrated SMOTE with ensemble-based machine learning for AQI classification, demonstrating that oversampling can improve predictive performance under imbalanced conditions. Similarly, [7] investigated boosting-based algorithms, including AdaBoost, XGBoost, and LightGBM, showing that while these models improve aggregate predictive performance, substantial disparities across classes persist, particularly for minority classes. Collectively, these studies demonstrate that resampling and ensemble-based approaches can improve predictive performance under imbalanced AQI classification scenarios. However, the reported improvements are predominantly assessed using aggregate evaluation metrics, with limited emphasis on the robustness of minority-class detection under extreme imbalance conditions. Consequently, a critical limitation may not solely reside in model selection itself, but also in the framework used for performance evaluation. In many existing studies, evaluation strategies rely heavily on global metrics such as accuracy, F1-score, and MCC, which may provide overly optimistic performance estimates under severe class imbalance because the majority class disproportionately influences the optimization objective. As a result, models optimized toward aggregate performance metrics may underrepresent minority-class patterns, particularly hazardous air quality conditions, potentially leading to an overestimation of real-world model reliability. In environmental risk contexts, such misclassification may delay public response and increase population exposure to dangerous pollutants [8]. Therefore, reliance on global metrics alone risks masking critical weaknesses in minority detection performance [9]. Furthermore, some approaches simplify imbalance issues by merging rare categories, which alters the original AQI categories hierarchy and reduces the granularity of risk interpretation. In addition, under extreme imbalance where class ratios exceed 1:100, hyperparameter optimization alone often fails to address the fundamental issue of minority class underrepresentation. Without a structured evaluation strategy that explicitly prioritizes minority detection and model stability, models may still struggle to reliably identify critical air quality conditions [10].

To address these limitations, this study proposes a new evaluation framework for multi-class AQI classification under extreme imbalance conditions (approximately 1:173). Unlike previous studies that often suffer from data leakage or prioritize aggregate metrics, this work adopts a rigorous, reliability-oriented architecture based on a Nested Cross-Validation procedure and an independent 20% hold-out test set to ensure unbiased assessment on unseen hazardous conditions. This framework systematically isolates the impact of hyperparameter optimization from oversampling techniques such as SMOTE and ADASYN through an Ablation Study approach. By prioritizing the Macro F1-Score and minority-class recall as primary safety indicators, the proposed framework aims to reduce the underrepresentation of hazardous air quality patterns specifically for the Unhealthy and Unhealthy for Sensitive Groups categories that are often masked by global accuracy. This study contributes a new reliability-oriented evaluation framework for AQI classification under extreme class imbalance conditions by integrating Nested Cross-Validation, leakage-aware validation procedures, minority-class focused evaluation metrics, and systematic ablation analysis to support more reliable decision-making in air quality monitoring systems.

2. RESEARCH METHOD

2.1 Research Design

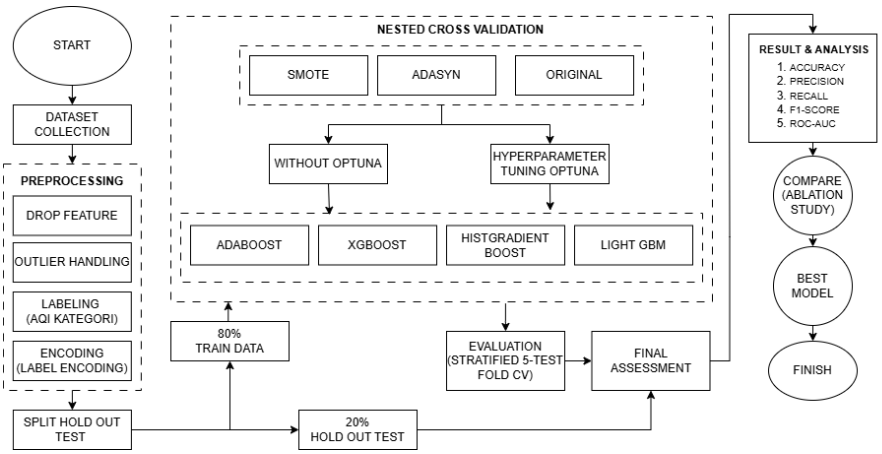

Figure 1. Model Workflow

.

As visualized in Figure 1, the experimental workflow begins with the "Global Air Quality 2025-6 Cities" dataset, which is subsequently divided into development and independent hold-out sets for model evaluation, this study proposes a new evaluation framework for multi-class AQI classification under extreme class imbalance (1:173), implementing a strict, reliability-oriented architecture that initially divides the dataset into an 80% development set and a 20% independent hold-out test set to act as a methodological firewall against data leakage. By isolating the 20% hold-out set from the model development phase, the framework ensures an unbiased assessment of generalization capabilities on unseen hazardous conditions, addressing the inherent limitations of conventional accuracy-driven evaluation in safety-critical monitoring [11]. Within this architecture, the 80% development set is processed using a Nested Cross-Validation procedure to conduct a systematic Ablation Study, comparing six distinct configurations, original data versus resampling strategies (SMOTE and ADASYN) and baseline algorithms versus tuned models optimized via Optuna targeting the Macro F1-Score. To maintain statistical integrity, all resampling and hyperparameter optimization processes are performed strictly within the internal training folds of the cross-validation, after which the best-performing model configuration is subjected to a Final Performance Assessment on the entirely unseen 20% hold-out test set to quantify reliability through the Macro F1-Score and minority-class recall.

2.2 Dataset Description

The dataset used in this study was obtained from the publicly available "Global Air Quality 2025-6 Cities" dataset hosted on Kaggle (Youssef Elebiary, 2025). It consists of 52,704 air quality observations collected at an hourly resolution from January to December 2025 (GMT time zone). The dataset includes six globally distributed urban locations, namely Brasilia, Cairo, Dubai, London, New York, and Sydney, representing diverse environmental and pollution conditions. Each record contains pollutant measurements, including carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), particulate matter (PM2.5), and PM10. These features are widely used indicators in standard AQI computation and represent critical environmental health parameters [12]. The dataset was selected due to its multi-city coverage, high temporal resolution, and comprehensive pollutant representation, enabling evaluation under heterogeneous real-world air quality conditions. Such variability reflects practical environmental monitoring scenarios where pollutant levels fluctuate across time and location, naturally introducing severe class imbalance. The class distribution is highly skewed. The Unhealthy AQI category accounts for only approximately 0.41% of the total dataset, resulting in an extreme imbalance ratio of approximately 1:173 relative to the majority class. This characteristic makes the dataset particularly suitable for evaluating reliability-oriented classification frameworks under extreme imbalance conditions. Such imbalance presents significant challenges for multi-class classification models, as learning algorithms tend to bias predictions toward majority classes. In safety-sensitive environmental monitoring systems, failure to detect hazardous AQI categories may lead to severe public health consequences [13]. Therefore, robust imbalance aware modeling strategies are required. The class distribution prior to any resampling strategy is illustrated in Figure 2.

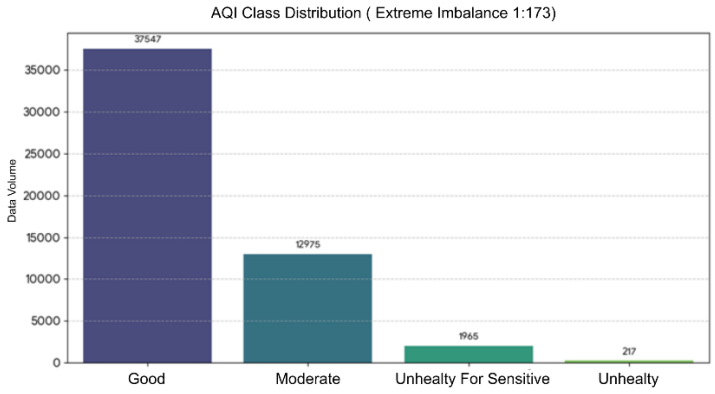

Figure 2. AQI Class Distribution

To provide a clearer overview of the dataset used in this study, the main characteristics and attributes are summarized in Table 1.

Table 1.Atribute Description

Feature	Data Type	Description
Date	Nominal	Date of data recording
City	Nominal	Name of the measurement location
CO	Numerical	Carbon monoxide concentration (ppm)
CO_2	Numerical	Carbon dioxide concentration (ppm)
NO_2	Numerical	Nitrogen dioxide concentration (ppb)
SO_2	Numerical	Sulfur dioxide concentration (ppb)
O_3	Numerical	Ozone concentration (ppb)
PM2.5	Numerical	Fine particulate matter < 2.5 μm (μg/m³)
PM10	Numerical	Coarse particulate matter < 10 μm (μg/m³)
AQI	Numerical	Air Quality Index

2.3 Data Preprocessing and Outlier Handling

Data preprocessing was conducted to improve numerical stability and prevent distortion caused by extreme pollutant values [14]. The CO_2 attribute was excluded due to a substantial proportion of missing entries. The Date and City attributes were removed as they do not directly contribute to numerical AQI classification and may introduce irrelevant categorical variation. An adaptive distribution-aware preprocessing strategy was implemented. Skewness for each numerical feature was computed using the Fisher Pearson coefficient. Features exhibiting high skewness (|skewness| > 1) were transformed using a logarithmic transformation to reduce distribution asymmetry and stabilize variance. For non-transformed features, extreme observations were handled using a conservative Interquartile Range (IQR) capping strategy [15]. IQR is defined as shown in Equation (1)-(3):

$$\text{IQR} = \text{Q3} - Q1 \quad (1)$$

$$\text{Lower Bound} = \text{Q1} - 3 \times IQR \quad (2)$$

$$\text{Upper Bound} = \text{Q3} + 3 \times IQR \quad (3)$$

Values outside these limits were capped (winsorized) rather than removed. The use of a 3 × IQR multiplier ensures that only extreme anomalies are constrained, preserving legitimate pollutant spikes that may represent real environmental events. All preprocessing operations, including skewness assessment, logarithmic transformation, IQR capping, and feature standardization, were performed exclusively within the internal training folds of the Nested Cross-Validation framework to prevent data leakage [16]. Feature scaling was applied using standardization to normalize feature magnitudes prior to model training. The AQI target variable was encoded into numerical labels for multi-class classification [17]. In addition, hyperparameter optimization was also conducted exclusively within the internal training folds of the Nested Cross-Validation framework, as detailed in Section 2.7.

2.4 SMOTE Oversampling

Class imbalance in the AQI dataset was addressed using the Synthetic Minority Over-sampling Technique (SMOTE). This method generates synthetic minority samples by interpolating between existing minority instances in the feature space. Instead of duplicating existing samples, SMOTE constructs new observations along the line segments connecting a minority instance and its nearest minority neighbors [18]. Synthetic samples were generated using the following formula shown in Equation (4):

$$x_{\text{new}} = x_i + \lambda(x_{nn} - x_i) \quad (4)$$

where x_i represents a minority class instance, x_{nn} denotes one of its nearest minority neighbors, and λ is a random interpolation factor within the interval [0,1]. Through this mechanism, SMOTE expands minority decision regions while preserving the local structure of the data distribution. In this study, SMOTE was applied within the training portion of each fold during Stratified 5-Fold Cross Validation to prevent data leakage. The oversampling process was configured to balance the dataset by increasing the number of samples in minority classes until they matched the size of the majority class.

2.5 ADASYN Oversampling

In addition to SMOTE, the Adaptive Synthetic Sampling (ADASYN) technique was employed to further address the extreme class imbalance present in the AQI dataset. ADASYN extends SMOTE by generating synthetic samples adaptively according to the learning difficulty of each minority instance [18]. Minority samples located in regions dominated by majority-class neighbors receive greater emphasis during

the oversampling process, enabling the classifier to focus more effectively on difficult decision boundaries. The local imbalance ratio for each minority instance is calculated as shown in Equation (5) [19]:

$$r_i = \frac{d_i}{N_i} \tag{5}$$

where r_i represents the local imbalance ratio for minority instance x_i, d_i denotes the number of majority-class neighbors surrounding the minority sample, and N_i represents the total number of nearest neighbors considered. A higher r_i value indicates that the minority sample is located in a more difficult classification region.
The normalized weight used to determine the proportion of synthetic samples generated for each minority instance is then computed as shown in Equation (6):

$$g_i = \frac{r_i}{\sum_{j=1}^{N_m} r_j} \tag{6}$$

where g_i denotes the normalized weight assigned to minority instance x_i, and N_m represents the total number of minority samples. Based on this weighting mechanism, minority instances located in more complex regions receive a larger proportion of synthetic sample generation.
The total number of synthetic samples to be generated is defined as shown in Equation (7):

$$G = \left(N_{maj} - N_{min}\right) \times \beta \tag{7}$$

where G denotes the total number of synthetic samples, N_{maj} represents the number of majority-class samples, N_{min} denotes the number of minority-class samples, and β is the oversampling coefficient controlling the desired balancing level.
Finally, synthetic samples are generated through linear interpolation between a minority instance and one of its minority nearest neighbors as follows in Equation (8):

$$x_{synth} = x_i + \delta \cdot (x_j - x_i) \tag{8}$$

where x_j is a randomly selected minority nearest neighbor of x_j, and δ (delta) is a random value within the interval [0,1]. Through this adaptive mechanism, ADASYN concentrates synthetic sample generation in sparse and difficult minority regions, thereby improving the classifier's sensitivity toward hazardous AQI categories under extreme imbalance conditions.

2.6 Boosting Model Architecture

This study evaluates four tree-based boosting algorithms, AdaBoost, XGBoost, HistGradientBoosting, and LightGBM. Each model was trained and evaluated independently, no model-level ensemble aggregation or voting mechanism was employed. The objective of this comparison is to examine the intrinsic robustness of each boosting framework under extreme class imbalance conditions in multi-class AQI classification. Boosting constructs a strong learner by sequentially combining weak learners to minimize prediction error [20]. At iteration m, the ensemble model is updated as shown in Equation (9):

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \tag{9}$$

where $h_m(x)$ represents the newly fitted weak learner and η denotes the learning rate controlling the contribution of each tree. This additive formulation enables gradual residual correction across iterations, allowing the model to refine decision boundaries progressively.

In gradient boosting frameworks, model optimization is performed by minimizing a differentiable loss function defined as shown in Equation (10):

$$\mathcal{L} = \sum_{i=1}^{n} l\left(y_i, F(x_i)\right) \tag{10}$$

where $l\left(y_i, F(x_i)\right)$ measures the discrepancy between the true label y_i and the predicted output $F(x_i)$

XGBoost extends this formulation by including an explicit regularization term to control model complexity, which can be shown in Equation (11):

$$L = \sum_{i=1}^{n} l\left(y_i, F(x_i)\right) + \Omega(f) \tag{11}$$

where $\Omega(f)$ penalizes model complexity to reduce overfitting and improve generalization.

HistGradientBoosting improves computational efficiency by discretizing continuous features into histogram bins prior to split evaluation. This histogram-based approach significantly accelerates training while maintaining predictive performance, making it suitable for structured environmental datasets of moderate size [21].

LightGBM further enhances scalability by adopting a leaf-wise tree growth strategy rather than traditional level-wise expansion. This allows deeper splits along high-gain branches and faster convergence, while regularization parameters constrain overfitting [22].

Each boosting model was configured using its native multi-class implementation (softmax / multinomial objective). No one-vs-rest (OvR) decomposition was applied. The loss functions used were as follows AdaBoost employs exponential loss, which increases weights for misclassified samples at each iteration XGBoost and LightGBM use multiclass log-loss to produce probabilistic predictions across AQI categories and HistGradientBoosting was configured with log-loss (deviance) for multi-class classification. This consistent use of probabilistic objectives (except for AdaBoost's exponential formulation) ensures comparable optimization behavior across models while preserving their intrinsic characteristics [23]. Given the extreme imbalance condition of the AQI dataset, boosting models are particularly suitable due to their iterative error-correction mechanism. When combined with imbalance aware preprocessing strategies, boosting enables progressive adaptation to minority class patterns while maintaining overall predictive stability [24].

2.7 Hyperparameter Optimization Using Optuna

Hyperparameter optimization was performed using Optuna, an automatic optimization framework based on the Tree-Structured Parzen Estimator (TPE) algorithm [25]. To address the extreme class imbalance and ensure a balanced trade-off between sensitivity and precision, the optimization objective was defined as the Macro Average F1-Score. The Macro F1-Score is the harmonic mean of Precision and Recall, calculated as follows in Equation (12):

$$\text{Macro F1} = \frac{1}{K}\sum_{i=1}^{K}\frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{12}$$

Where K represents the number of classes. Unlike global accuracy, the Macro F1-Score assigns equal weight to each category, preventing the majority classes from over-optimizing the model at the expense of the minority Unhealthy class.

The tuning process was integrated within the inner loop of the Nested Cross-Validation architecture (as shown in Figure 1). Optimization was executed independently for each data configuration, the original, SMOTE, and ADASYN distributions. This ensures that the resulting hyperparameters are specifically tuned to the unique data density of each scenario. During the optimization, Optuna conducted 50 trials per configuration to identify the parameter set that maximizes the Macro F1-Score. Once the optimal configuration was identified within the internal training folds, the model was evaluated on the unseen outer validation fold. This systematic approach allows for a transparent Ablation Study, demonstrating the combined impact of resampling techniques and targeted hyperparameter tuning on model robustness. The hyperparameter search space for each boosting model is summarized in Table 2.

Table 2.Hyperparameter search space for Optuna optimization.

Model	Hyperparameter	Search Range	Scale
AdaBoost	n_estimators	50 - 500	Integer
	learning_rate	0.01 - 1.0	Float
XGBoost	n_estimators	100 - 800	Integer
	learning_rate	0.01 - 0.3	Float
	max_depth	3 - 10	Integer
HistGradientBoosting	max_iter	100 - 800	Integer
	learning_rate	0.01 - 0.3	Log
	max_depth	3 - 12	Integer
LightGBM	n_estimators	100 - 800	Integer
	learning_rate	0.01 - 0.3	Log
	num_leaves	20 - 200	Integer

.

2.8 Validation Strategy

To obtain reliable performance estimates under the extreme 1:173 imbalance ratio, this study implements a multi-layered validation strategy centered on an 80% development set and a 20% independent hold-out set. The 20% hold-out set serves as a methodological firewall, remaining entirely isolated from all development phases including preprocessing parameter estimation, resampling, and hyperparameter tuning to provide a definitive assessment of the model's generalization on entirely unseen hazardous conditions. Within the development set, a nested cross-validation architecture is employed to decouple hyperparameter optimization from performance estimation, ensuring a bias-free evaluation process. The outer loop utilizes a stratified 5-fold cross-validation to preserve the original class distribution in every partition, while the inner loop conducts Optuna-based tuning and resampling exclusively within the internal training folds. SMOTE and ADASYN are specifically integrated into this inner training phase because they are mathematically suited for extreme skewness where minority signals are often discarded as statistical noise. SMOTE counteracts the majority dominance trap by expanding the minority class decision regions via linear interpolation [26], whereas ADASYN adaptively generates synthetic samples in low-density, hard-to-learn areas where the Unhealthy class overlaps with majority classes. Crucially, while training is conducted on these resampled distributions, the validation subsets in every fold always retain the original, unmodified 1:173 class distribution. This rigorous procedure prevents data leakage and ensures that the final performance metrics, reported as mean ± standard deviation, reflect real-world stability and the model's genuine discriminative power rather than over-optimistic artifacts resulting from artificial balancing.

2.9 Evaluation Metrics

Model performance was evaluated using metrics that reflect both overall predictive capability and safety-critical detection under extreme class imbalance. While global Accuracy is reported, it is not used as the primary benchmark because it tends to be over-optimistic in imbalanced contexts. The formulas for the evaluation metrics used in this study are presented in Equations (13) - (16).

Accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

For multi-class evaluation, class-wise precision and recall are computed as:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \tag{14}$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \tag{15}$$

The F1-Score for class i defined as:

$$F1_i = 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{16}$$

To account for the extreme 1:173 imbalance, this study utilizes Macro Averaging rather than weighted metrics. Unlike weighted averaging, which biases the result toward the majority class, Macro Average assigns equal importance to each AQI category, ensuring that poor performance on the minority Unhealthy class is heavily penalized [27]. The Macro F1-Score, which serves as the primary evaluation metric and the objective function during hyperparameter optimization, is defined as shown in Equation (17):

$$\text{Macro F1} = \frac{1}{\text{K}} \sum_{\text{i}=1}^{\text{K}} \frac{2 \times \text{Precision}_\text{i} \times \text{Recall}_\text{i}}{\text{Precision}_\text{i} + \text{Recall}_\text{i}} \tag{17}$$

Where K denotes the number of AQI classes. Given the safety-oriented objective of this study, special emphasis is placed on Minority Recall (Recall for the Unhealthy class). Minority Recall serves as the primary safety indicator, as false negatives in hazardous air quality detection categorizing dangerous air as safe may result in severe public health consequences. Finally, the Area Under the ROC Curve (AUC-ROC) is calculated using a One-vs-Rest (OvR) scheme to measure the model's discriminative power across all AQI categories. This evaluation framework ensures that model selection prioritizes balanced robustness and reliable detection under extreme imbalance conditions.

2.10 Model Selection

To provide an objective and quantitative basis for model selection, this study implements a systematic Ablation Study. This approach replaces subjective selection criteria with a rigorous decomposition of the proposed framework, evaluating the individual and combined contributions of hyperparameter optimization and resampling techniques [28]. The experimental design is structured into six distinct scenarios, as summarized in Table 3.

Table 3. Systematic Ablation Study Scenarios

Scenario	Data Configuration	Model Configuration	Primary Objective
S1 (Baseline)	Original (Imbalanced)	Default Hyperparameters	To establish the baseline performance without any interventions.
S2 (Resampling-S)	SMOTE	Default Hyperparameters	To isolate the performance impact of SMOTE oversampling.
S3 (Resampling-A)	ADASYN	Default Hyperparameters	To isolate the performance impact of ADASYN oversampling.
S4 (Tuning-O)	Original (Imbalanced)	Tuned (Optuna)	To isolate the impact of hyperparameter optimization via Optuna.
S5 (Hybrid-S)	SMOTE	Tuned (Optuna)	To evaluate the synergy between SMOTE and hyperparameter tuning.
S6 (Hybrid-A)	ADASYN	Tuned (Optuna)	To evaluate the synergy between ADASYN and hyperparameter tuning.

The evaluation of these scenarios follows a strict Macro F1-Score maximization priority. This strategy ensures that the selection process is driven by the model's ability to balance Precision and Recall across all classes, particularly the minority Unhealthy class, without relying on arbitrary thresholds or global accuracy. By comparing the results across these scenarios, the framework identifies the specific component responsible for performance gains, ensuring a transparent and evidence-based model selection. Model selection is strictly driven by the peak Macro F1-Score, providing an automated and quantitative solution to the Precision-Recall trade-off inherent in extreme class imbalance (1:173).

3. RESULTS AND DISCUSSION

3.1 Dataset Distribution Analysis

The distribution of AQI categories in the original dataset is highly imbalanced. Most observations belong to the Good and Moderate categories, whereas the Unhealthy class appears only in a very small number of samples. This condition produces an extreme imbalance ratio of approximately 1:173, which can bias classification models toward majority classes and reduce the ability to detect hazardous air quality conditions. To address this issue, two oversampling techniques were applied, namely SMOTE and ADASYN. These techniques generate synthetic samples for minority classes within the training data so that the representation of each category becomes more balanced. In this study, oversampling was performed until the number of samples in each AQI category reached the same level as the majority class, a transformation that is visually compared and illustrated in Figure 3.

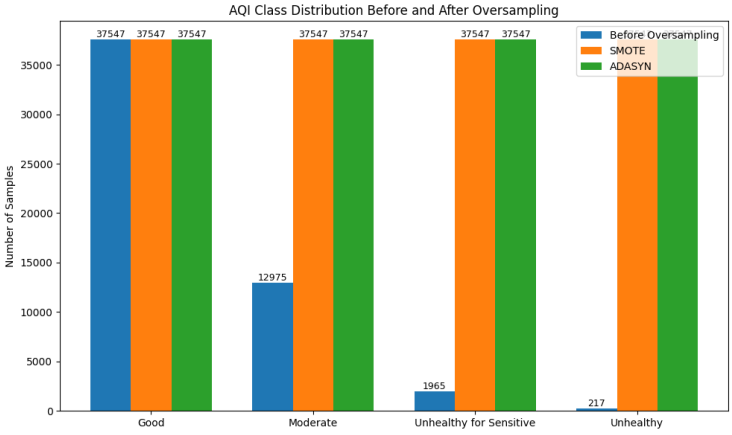

Figure 3. Class distribution before and after oversampling

3.2 Experimental Setup

The experiments were conducted to evaluate the robustness of boosting-based classification models under extreme class imbalance conditions using a systematic multi-stage framework. The dataset was initially partitioned into a stratified 80% development set and a 20% independent hold-out test set to ensure a final,

.

unbiased performance verification. For the model development phase, four ensemble boosting algorithms were considered AdaBoost, XGBoost, HistGradientBoosting, and LightGBM. Performance estimation and model selection were performed using a stratified 5-fold Nested Cross-Validation (NCV) architecture. Within the outer loop of this architecture, the development set was divided into five folds to assess generalization stability. The inner loop was utilized for hyperparameter optimization via Optuna across 50 trials for each scenario. To maintain methodological integrity, imbalance handling techniques specifically SMOTE and ADASYN were applied exclusively to the training portion within each internal fold. This procedure ensures that the models were trained on synthetically balanced distributions while being validated on the original, unmodified class distribution, thereby preventing data leakage, and reflecting real-world performance. Furthermore, the experiments were structured into six ablation scenarios (S1-S6) to isolate the individual and combined effects of resampling and hyperparameter tuning on the Macro F1-Score and minority class detection.

3.3 Model and Optimization Configuration

Hyperparameter optimization was conducted to align model complexity with the specific characteristics of the imbalanced AQI dataset. To eliminate the risk of data leakage during hyperparameter optimization, this study distinguishes between the tuning phase and the final evaluation phase through a Nested Cross-Validation procedure. The performance metrics reported in this section are derived from the outer loop (for cross-validation results) and an independent 20% hold-out test set (for final unbiased verification). This architecture ensures that the data used to guide Optuna's search in the inner loop remains strictly isolated from the data used to report the final model reliability. The resulting configurations, obtained after 50 trials per scenario using Optuna, demonstrate a significant departure from baseline settings. These adjustments indicate that the optimization process prioritized more granular learning and increased ensemble depth to capture the rare patterns of the Unhealthy class. Table 4 summarizes the baseline parameters alongside the optimal values identified for the original, SMOTE, and ADASYN distributions.

Table 4. Baseline and tuned hyperparameter configurations obtained from Optuna optimization.

Model	Hyperparameter	Baseline	Tuned Original	Tuned SMOTE	Tuned ADASYN
AdaBoost	n_estimators	50	359	353	110
	learning_rate	1	0.1075	0.7141	0.1083
XGBoost	n_estimators	100	431	597	438
	learning_rate	0.3	0.0573	0.0642	0.1408
	max_depth	6	10	8	9
HistGradientBoosting	max_iter	100	240	761	712
	learning_rate	0.1	0.0467	0.1447	0.123
	max_depth	None	11	10	9
LightGBM	n_estimators	100	748	643	610
	learning_rate	0.1	0.1026	0.0249	0.098
	num_leaves	31	193	172	102

The optimized configurations reveal consistent trends in model adaptation across the evaluation scenarios. In most cases, the tuned models employ a significantly larger number of boosting iterations or estimators compared to the baseline, suggesting that standard settings are insufficient for resolving complex class boundaries under extreme imbalance. For instance, the LightGBM model in the Hybrid-S (S5) scenario utilizes 643 estimators with a substantially reduced learning rate of 0.0249. This combination facilitates a more gradual and precise error-correction process, allowing the model to refine its decision regions for minority samples without overfitting the majority Good class. Furthermore, the structural complexity of gradient-based models increased, as evidenced by the deeper tree structures in XGBoost and HistGradientBoosting, as well as the expanded leaf capacity in LightGBM. These adjustments provide the models with the necessary expressive power to capture intricate feature interactions that characterize sudden pollutant spikes. The divergence in parameters between the SMOTE and ADASYN scenarios also proves that the optimization process is highly sensitive to the specific density of synthetic data, confirming that a universal parameter configuration is inadequate for robust air quality monitoring.

3.4 Comparative Performance Analysis across Ablation Scenarios

This section provides a systematic evaluation of four ensemble boosting algorithms across six experimental scenarios (S1-S6). The analysis deconstructs the individual and synergistic impacts of original data distributions, oversampling techniques (SMOTE and ADASYN), and Optuna-based hyperparameter optimization on model robustness under an extreme 1:173 class imbalance. All reported metrics utilize Macro-averaging to ensure that the detection of the minority Unhealthy class is penalized equally with majority classes, preventing misleadingly high global accuracy from masking systemic failures.

3.4.1. Cross-Model Best Performance Comparisons

Table 5 summarizes the peak performance reached by each boosting algorithm across the experimental scenarios. While global metrics provide an overview of model capability, a more granular evaluation is required to ensure that the high performance is consistent across all Air Quality Index (AQI) risk levels, especially under the extreme 1:173 class imbalance.

Table 5. Best Performance Comparison Across Algorithms (Mean of 5-Fold CV)

Model	Best Config	Accuracy	Macro Prec	Macro Rec	Macro F1	ROC-AUC
AdaBoost	S1 (Baseline)	0.7889	0.5187	0.5554	0.4909	0.8581
XGBoost	S5 (Hybrid-S)	0.9151	0.7427	0.8105	0.7725	0.9832
HistGradientBoost	S5 (Hybrid-S)	0.9128	0.7467	0.8070	0.7737	0.9812
LightGBM	**S5 (Hybrid-S)**	**0.9187**	**0.7541**	**0.8052**	**0.7772**	**0.9835**

To address the requirement for per-class reliability and to validate the safety-oriented nature of the proposed framework, Table 6 deconstructs the F1-Score of the best-performing model (LightGBM) across each AQI category. This comparison highlights how the S5 configuration bridges the performance gap in minority classes that are typically neglected by baseline models.

Table 6. Class-wise F1-Score Comparison for LightGBM: Baseline (S1) vs. Proposed (S5)

AQI Category	Baseline (S1)	Proposed (S5)	Improvement
Good	0.9574	0.9573	-0.01%
Moderate	0.8349	0.8428	0.95%
Unhealthy for Sensitive Groups	0.4261	0.5627	32.08%
Unhealthy	0.7235	0.7460	3.10%

The cross-model comparison in Table 5 establishes LightGBM (S5) as the most robust architecture, achieving a Macro F1-Score of 0.7772. The class-wise breakdown in Table 6 reveals that the framework's primary strength lies in its ability to enhance minority class detection without degrading majority class precision. The most significant improvement is observed in the Unhealthy for Sensitive Groups category, where the F1-score increased by 32.08% (from 0.4261 to 0.5627). This demonstrates that the integration of SMOTE and Optuna tuning successfully captured the nuanced patterns of moderate-to-high pollution levels.

Furthermore, the critical Unhealthy class maintains a strong F1-score of 0.7460 in the S5 configuration. While the baseline (S1) already showed competitive performance for this class, the proposed hybrid approach further refined the decision boundaries, yielding a 3.10% increase. Crucially, these gains were achieved with a negligible change (-0.01%) in the Good category, proving that the model successfully reconfigures its predictive bias to be more risk-aware. This dual stability across the entire AQI spectrum ensures that the proposed framework is not only accurate but also reliable for safety-critical public health warnings.

The findings of this study are consistent with previous AQI classification research demonstrating the effectiveness of resampling and boosting-based approaches under imbalanced conditions. In particular, [6] demonstrated that integrating SMOTE with ensemble-based machine learning algorithms improved AQI classification performance under imbalanced conditions, highlighting the importance of oversampling strategies for minority representation. Similarly, [7] reported a Macro F1-Score of 0.52 in PM2.5 classification using boosting-based algorithms under imbalanced conditions, with class-wise F1-scores of 0.34 and 0.59 for the Moderate and Unhealthy categories, respectively.

However, unlike previous studies that primarily focused on localized environmental settings and aggregate predictive performance, the proposed framework in this study was evaluated using a multi-city global dataset consisting of six urban regions with diverse environmental characteristics and an extreme imbalance ratio of approximately 1:173. Despite these more challenging conditions, the LightGBM Hybrid-SMOTE (S5) configuration achieved a Macro F1-Score of 0.8079 on the independent hold-out test set while improving Recall for the critical Unhealthy category from 0.4651 to 0.6512. Furthermore, the proposed framework incorporates a reliability-oriented evaluation architecture through Nested Cross-Validation, independent hold-out testing, minority-class focused evaluation metrics, and systematic ablation analysis. These findings suggest that reliability-oriented evaluation strategies may provide a more dependable framework for hazardous AQI detection under heterogeneous environmental conditions.

3.4.2. Detailed Ablation Analysis of Selected Model (LightGBM)

To isolate the technical drivers of performance gains, Table 7 presents the ablation breakdown for the LightGBM model. This systematic decomposition highlights the necessity of the hybrid approach over standalone interventions.

.

Table 7. Performance Breakdown for LightGBM Ablation Study (Mean of 5-Fold CV)

Scenario	Data Config	Model Config	Acc	M-Prec	M-Rec	M-F1	AUC
S1 (Baseline)	Original	Default	0.9177	0.7789	0.7094	0.7355	0.9829
S2 (Resampling-S)	SMOTE	Default	0.9009	0.709	0.8121	0.7519	0.9812
S3 (Resampling-A)	ADASYN	Default	0.8710	0.6568	0.8119	0.7145	0.9765
S4 (Tuning-O)	Original	Tuned	0.9235	0.8111	0.7271	0.7577	0.9830
S5 (Hybrid-S)	**SMOTE**	**Tuned**	**0.9187**	**0.7541**	**0.8052**	**0.7772**	**0.9835**
S6 (Hybrid-A)	ADASYN	Tuned	0.9140	0.7418	0.7983	0.7674	0.9840

The experimental data reveals that the synergy achieved in the hybrid scenario (S5) represents the most significant methodological breakthrough of this study. While standalone interventions such as oversampling (S2) or hyperparameter tuning (S4) yielded marginal improvements, only the integration of SMOTE with Optuna-based optimization successfully resolved the structural bias toward the majority class. This jump in LightGBM's Macro F1-score from the baseline 0.7355 (S1) to 0.7772 (S5) proves that a balanced dataset is fundamentally limited without a precise parameter recipe just as optimal parameters cannot reach peak performance without a sufficient minority class signal to learn from. Furthermore, the failure of default parameters in the baseline scenario (S1) highlights a critical blind spot in conventional environmental monitoring, where reliance on factory settings often leads to models that systematically overlook hazardous conditions despite maintaining high global accuracy.

A deeper examination of the trade-offs between precision and recall further justifies the selection of the S5 configuration. Standalone oversampling techniques in S2 and S3 significantly increased sensitivity, yet they severely compromised macro precision by introducing synthetic noise that led to excessive false alarms. The S5 scenario corrected this imbalance through targeted hyperparameter tuning, which refined the model's decision boundaries and stabilized precision without sacrificing the detection of minority hazardous events. Additionally, the comparative stability of SMOTE-based scenarios over ADASYN suggests that for this specific urban air quality feature space, linear interpolation between minority neighbors provided a more effective representation for LightGBM's leaf-wise tree structure than density-based generation.

Finally, the integrity of these findings is underpinned by the rigorous validation architecture employed in this study. By utilizing a Nested Cross-Validation procedure, the reported Macro F1-score of 0.7772 serves as a realistic and reliable estimate of the model's generalization capabilities. This methodological firewall ensures that the performance gains are not merely artifacts of data leakage or artificial balancing, but rather a genuine reflection of the model's discriminative power. Consequently, prioritizing the Macro F1-score ensures that model selection remains safety-oriented, penalizing systemic failures in minority class detection that global metrics often mask.

3.5 Safety-Critical Evaluation: Minority Class Detection and Robustness

This section evaluates the effectiveness of the proposed framework by dissecting the model's performance on categories with a low frequency of occurrence but a significant impact on the final classification results. The primary focus of this analysis is to compare the sensitivity of the baseline model (S1) with the hybrid configuration (S5) in detecting air conditions within the Unhealthy for Sensitive Groups and Unhealthy categories.

Table 8. Minority Class Performance Metrics Comparison (Mean 5-Fold CV)

Metric	Baseline (S1)	Proposed (S5)	Improvement
Recall (Unhealthy)	0.3393	0.6267	+0.2874
F1-Score (Unhealthy for Sensitive Groups)	0.4261	0.5627	+0.1366
F1-Score (Unhealthy)	0.7235	0.7460	+0.0225
Standard Deviation (Unhealthy Recall)	± 0.038	± 0.014	-0.0240

The data in Table 8 indicates a shift in the model's predictive characteristics following the implementation of SMOTE integration and hyperparameter optimization. The most significant finding is observed in the Recall metric for the Unhealthy category, which increased from 0.3393 to 0.6267. Technically, this indicates that the hybrid model is capable of identifying more positive samples in hazardous air conditions compared to the standard model, which previously failed to capture approximately 66% of cases in this category.

In the Unhealthy for Sensitive Groups category, the model experienced an increase in F1-Score of 0.1366. Based on the baseline data, this category was the weakest point in the classification, with the lowest F1-Score (0.4261). The implementation of the S5 scenario proved capable of improving precision and sensitivity in this transitional category, although the improvement was not as high as that observed in the majority categories. It is important to note that the improved detection capability for minority classes was accompanied by a decrease in standard deviation from ± 0.038 to ± 0.014. This confirms that the hybrid model not only enhances sensitivity toward minority classes but also exhibits more consistent predictive stability

across cross-validation folds. This stability indicates that the adjustment of decision boundaries through resampling and tuning techniques does not compromise model robustness when encountering data variations.

3.6 Final Model Justification

Based on the multi-stage evaluation across various ablation scenarios, the LightGBM configuration integrated with SMOTE and Optuna-tuned hyperparameters (S5) is selected as the final model. This selection is justified by the model's superior ability to balance global metric stability with high sensitivity toward hazardous air quality categories. To ensure the model's generalizability and to guard against overfitting, a final validation was conducted using an independent hold-out test set (comprising 20% of the total data).

Table 9. Performance Comparison on Independent Hold-out Test Set (S1 vs. S5)

Metric	Baseline (S1)	Proposed (S5)	Improvement
Accuracy	0.9244	0.9280	0.39%
Macro Precision	0.8371	0.7923	-5.35%
Macro Recall	0.7484	0.8249	10.22%
Macro F1-Score	0.7832	0.8079	3.15%
ROC-AUC	0.9830	0.9847	0.17%

The results presented in Table 9 demonstrate that the LightGBM S5 model maintains stable performance when encountered with unseen data patterns. The Macro F1-Score of 0.8079 on the hold-out set is slightly higher than the 5-fold cross-validation mean (0.7772), confirming that the model does not suffer from overfitting and instead exhibits robust predictive power. The primary justification for the proposed framework lies in the Unhealthy Recall metric, which reached 0.6512, representing a 39.99% improvement over the baseline model (0.4651). Additionally, the model maintains a consistently high ROC-AUC of 0.9847, indicating exceptional discriminative power across all AQI classes. While a slight decrease in Macro Precision was observed, this is interpreted as a necessary technical trade-off to enhance minority class sensitivity. In the context of air quality monitoring, the ability to accurately detect hazardous conditions is significantly more critical than maintaining high global precision if it results in the failure to identify public health risks. The consistency of these results across both cross-validation and independent testing confirms that the S5 configuration provides the most reliable foundation for safety-oriented environmental monitoring systems.

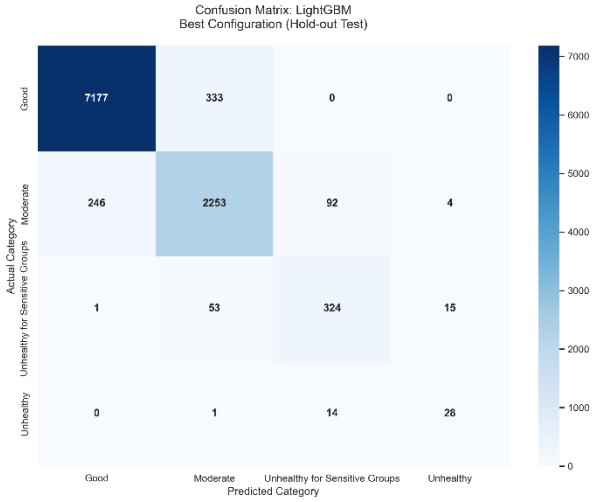

Figure 4. Confusion Matrix of Selected LightGBM Configuration

A more granular analysis of the model's predictive behavior is illustrated through the Confusion Matrix on the independent hold-out set (Figure 4). While global metrics confirm the model's overall reliability, analyzing the error profile is critical to validating the framework's safety-oriented objective. The matrix reveals how the LightGBM S5 model distributes its classification errors across the Air Quality Index (AQI) spectrum. The empirical data demonstrates a robust True Positive detection for the majority classes, correctly classifying 7,177 out of 7,510 Good samples and 2,253 out of 2,595 Moderate samples. However, the primary justification for this model lies in its handling of the minority risk categories. For the critical Unhealthy class (43 total instances), the model successfully identified 28 instances correctly. Crucially, out of the 15 misclassifications in this category, 14 instances were classified into the adjacent Unhealthy for Sensitive Groups category. Only a single instance was misclassified as Moderate, and zero instances were catastrophically predicted as Good. This error distribution proves that the model predominantly generates soft errors. Even when the model fails

.

to predict the exact peak of a hazardous event, 97.6% (42 out of 43) of the Unhealthy conditions are still flagged within the heightened risk and warning zones (Unhealthy for Sensitive Groups or Unhealthy).

Furthermore, the model exhibited exceptional capability in the transitional Unhealthy for Sensitive Groups category, correctly identifying 324 out of 393 instances. This high recall in the intermediate hazard zone ensures that early degradation in air quality is captured before reaching critical levels. By systematically confining misclassifications to adjacent risk categories and minimizing drastic false negatives, the Hybrid-SMOTE decision boundaries successfully enforce a conservative, risk-aware predictive mechanism suitable for public health early warning deployments.

4. CONCLUSION

This study proposed a safety-oriented machine learning framework to address extreme class imbalance (1:173) in Air Quality Index (AQI) classification. By integrating SMOTE-based resampling with Optuna hyperparameter tuning, the framework improved minority-class detection while maintaining stable overall predictive performance. Experimental results identified the LightGBM Hybrid-SMOTE (S5) configuration as the most robust model, achieving a Macro F1-Score of 0.8079 on the independent hold-out set. The proposed approach also improved Recall for the critical Unhealthy category from 0.4651 to 0.6512 compared to the baseline model. The confusion matrix analysis showed that most prediction errors remained within neighboring AQI risk categories, thereby reducing severe false-negative predictions for hazardous air quality conditions. These findings demonstrate that reliability-oriented evaluation strategies based on Macro F1-Score and minority-class recall provide a more reliable evaluation approach than accuracy-driven assessment under extreme class imbalance conditions.

Despite these promising results, several limitations should be acknowledged. Although this study incorporates more geographically diverse urban regions compared to previous localized AQI studies, the dataset is still limited to six cities and may not fully capture the environmental variability of other regions with different pollution characteristics and meteorological patterns. In addition, the proposed framework relies primarily on pollutant concentration features, while excluding other contextual variables such as meteorological dynamics, traffic density, and industrial activity that may further influence AQI behavior. Furthermore, although SMOTE-based oversampling improved minority class detection, synthetic interpolation in extremely sparse minority regions may still introduce sensitivity to noise and boundary distortion. Therefore, future research should investigate more adaptive imbalance-learning strategies, external cross regional validation, and real time streaming implementations to further evaluate the robustness and generalizability of safety-oriented AQI classification frameworks.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict on interest.

REFERENCES

[1] O. Sokolova, A. Yurgenson, and V. Shakhov, “Development of Air Quality Monitoring Systems: Balancing Infrastructure Investment and User Satisfaction Policies,” *Sensors*, vol. 25, no. 3, p. 875, Jan. 2025, doi: 10.3390/s25030875.

[2] Z. Zhong *et al.*, “Dynamic associations between long-term exposure to ambient air pollution and respiratory-cardiovascular diseases: A trajectory analysis of a prospective study,” *Ecotoxicol. Environ. Saf.*, vol. 306, p. 119329, Nov. 2025, doi: 10.1016/j.ecoenv.2025.119329.

[3] M. M. Abdelmalek, H. Mahmoud, and H. Shokry, “Prognosis of air quality index and air pollution using machine learning techniques,” *Sci. Rep.*, vol. 15, no. 1, p. 25890, Jul. 2025, doi: 10.1038/s41598-025-11260-y.

[4] V. R. Folifack Signing *et al.*, “IoT-based monitoring system and air quality prediction using machine learning for a healthy environment in Cameroon,” *Environ. Monit. Assess.*, vol. 196, no. 7, p. 621, Jul. 2024, doi: 10.1007/s10661-024-12789-7.

[5] R. S. Rao, L. R. Kalabarige, B. Alankar, and A. K. Sahu, “Multimodal imputation-based stacked ensemble for prediction and classification of air quality index in Indian cities,” *Computers and Electrical Engineering*, vol. 114, p. 109098, Mar. 2024, doi: 10.1016/j.compeleceng.2024.109098.

[6] A. J. Barid, H. Hadiyanto, and A. Wibowo, “Optimization of the algorithms use ensemble and synthetic minority oversampling technique for air quality classification,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 3, p. 1632, Mar. 2024, doi: 10.11591/ijeecs.v33.i3.pp1632-1640.

[7] T. Toharudin *et al.*, “Boosting Algorithm to Handle Unbalanced Classification of PM2.5Concentration 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.

[8] I. Tanasa, M. Cazacu, and B. Sluser, “Air Quality Integrated Assessment: Environmental Impacts, Risks and Human Health Hazards,” *Applied Sciences*, vol. 13, no. 2, p. 1222, Jan. 2023, doi: 10.3390/app13021222.

[9] N. Masseran, M. A. M. Safari, and R. R. M. Tajuddin, “Probabilistic classification of the severity classes of unhealthy air pollution events,” *Environ. Monit. Assess.*, vol. 196, no. 6, p. 523, Jun. 2024, doi: 10.1007/s10661-024-12700-4.

[10] G. Velarde *et al.*, “Tree boosting methods for balanced and imbalanced classification and their robustness over time in risk assessment,” *Intelligent Systems with Applications*, vol. 22, p. 200354, Jun. 2024, doi: 10.1016/j.iswa.2024.200354.

[11] S. Ketu and P. K. Mishra, “Scalable kernel-based SVM classification algorithm on imbalance air quality data for proficient healthcare,” *Complex & Intelligent Systems*, vol. 7, no. 5, pp. 2597–2615, Oct. 2021, doi: 10.1007/s40747-021-00435-5.

[12] Z.-Y. Chen, H. Petetin, R. F. Méndez Turrubiates, H. Achebak, C. Pérez García-Pando, and J. Ballester, "Population exposure to multiple air pollutants and its compound episodes in Europe," *Nat. Commun.*, vol. 15, no. 1, p. 2094, Mar. 2024, doi: 10.1038/s41467-024-46103-3.

[13] W. Chandra, B. Suprihatin, and Y. Resti, "Median-KNN Regressor-SMOTE-Tomek Links for Handling Missing and Imbalanced Data in Air Quality Prediction," *Symmetry (Basel).*, vol. 15, no. 4, p. 887, Apr. 2023, doi: 10.3390/sym15040887.

[14] H. Alkabbani, A. Ramadan, Q. Zhu, and A. Elkamel, "An Improved Air Quality Index Machine Learning-Based Forecasting with Multivariate Data Imputation Approach," *Atmosphere (Basel).*, vol. 13, no. 7, p. 1144, Jul. 2022, doi: 10.3390/atmos13071144.

[15] M. Karmoude *et al.*, "Machine learning for air quality prediction and data analysis: Review on recent advancements, challenges, and outlooks," *Science of The Total Environment*, vol. 1002, p. 180593, Nov. 2025, doi: 10.1016/j.scitotenv.2025.180593.

[16] S. P. Praveen *et al.*, "Enhanced feature selection and ensemble learning for cardiovascular disease prediction: hybrid GOL2-2 T and adaptive boosted decision fusion with babysitting refinement," *Front. Med. (Lausanne).*, vol. 11, Jul. 2024, doi: 10.3389/fmed.2024.1407376.

[17] K. Abnoosian, R. Farnoosh, and M. H. Behzadi, "Prediction of diabetes disease using an ensemble of machine learning multi-classifier models," *BMC Bioinformatics*, vol. 24, no. 1, p. 337, Sep. 2023, doi: 10.1186/s12859-023-05465-z.

[18] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, p. 87, Jun. 2024, doi: 10.1186/s40537-024-00943-4.

[19] Haibo He, Yang Bai, E. A. Garcia, and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, IEEE, Jun. 2008, pp. 1322–1328. doi: 10.1109/IJCNN.2008.4633969.

[20] Y.-W. Wang, H.-C. Yang, Y.-H. Chen, and C.-Y. Guo, "Generalized Estimating Equations Boosting (GEEB) machine for correlated data," *J. Big Data*, vol. 11, no. 1, p. 20, Jan. 2024, doi: 10.1186/s40537-023-00875-5.

[21] S. Saminathan and C. Malathy, "Ensemble-based classification approach for PM2.5 concentration forecasting using meteorological data," *Front. Big Data*, vol. 6, Jun. 2023, doi: 10.3389/fdata.2023.1175259.

[22] X. Zhang, X. Jiang, and Y. Li, "Prediction of air quality index based on the SSA-BiLSTM-LightGBM model," *Sci. Rep.*, vol. 13, no. 1, p. 5550, Apr. 2023, doi: 10.1038/s41598-023-32775-2.

[23] J. Ugirumurera, E. A. Bensen, J. Severino, and J. Sanyal, "Addressing bias in bagging and boosting regression models," *Sci. Rep.*, vol. 14, no. 1, p. 18452, Aug. 2024, doi: 10.1038/s41598-024-68907-5.

[24] M. S. Sawah, H. Elmannai, A. A. El-Bary, Kh. Lotfy, and O. E. Sheta, "Improving air quality prediction using hybrid BPSO with BWAO for feature selection and hyperparameters optimization," *Sci. Rep.*, vol. 15, no. 1, p. 13176, Apr. 2025, doi: 10.1038/s41598-025-95983-y.

[25] H. Shao, X. Liu, D. Zong, and Q. Song, "Optimization of diabetes prediction methods based on combinatorial balancing algorithm," *Nutr. Diabetes*, vol. 14, no. 1, p. 63, Aug. 2024, doi: 10.1038/s41387-024-00324-z.

[26] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4903–4923, Jul. 2024, doi: 10.1007/s10994-022-06296-4.

[27] S. Farhadpour, T. A. Warner, and A. E. Maxwell, "Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices," *Remote Sens. (Basel).*, vol. 16, no. 3, p. 533, Jan. 2024, doi: 10.3390/rs16030533.

[28] S. Sheikholeslami, M. Meister, T. Wang, A. H. Payberah, V. Vlassov, and J. Dowling, "AutoAblation: Automated Parallel Ablation Studies for Deep Learning," in *Proceedings of the 1st Workshop on Machine Learning and Systems*, New York, NY, USA: ACM, Apr. 2021, pp. 55–61. doi: 10.1145/3437984.3458834.

.

BIOGRAPHIES OF AUTHORS

Made Yudi Dwipayana, S.Kom is currently pursuing a Master's degree in Information Systems at ITB STIKOM Bali, Indonesia. He received his bachelor's degree in Computer Systems from STIKOM Bali. His research interests include machine learning, data mining, and data analytics, particularly in environmental data analysis. His current research focuses on safety-oriented air quality index classification using optimized boosting models and oversampling techniques for imbalanced datasets. He can be contacted at: yudidwipayana@stikom-bali.ac.id

Dr. Gede Angga Pradipta, S.T., M.Eng received the bachelor's degree in computer informatics from Universitas Atma Jaya (UAJY), Yogyakarta, Indonesia, in 2012, and the master's degree in information technology from Universitas Gadjah Mada (UGM), Yogyakarta, in 2014, where he is currently pursuing the Ph.D. degree with the Department of Computer Science and Electronics, Faculty of Natural Sciences. His research interests include machine learning, imbalanced data learning, and image processing. He can be contacted at: angga_pradipta@stikom-bali.ac.id

Dr. Dandy Pramana Hostiadi, S.Kom., M.T. Dandy Pramana Hostiadi (Member, IEEE) was born in Surabaya, Indonesia. He received a bachelor's degree in computer systems from STMIK STIKOM Bali, a master's degree in electrical engineering, focusing on information systems and computer management from Udayana University, and a Ph.D. degree from the Institut Teknologi Sepuluh Nopember (ITS), where he worked on finding ways to detect botnet activities in computer networks. In addition to his research, he was an Associate Professor at the Institut Teknologi dan Bisnis STIKOM Bali, where he shared his knowledge of network security and forensics with students. He is dedicated to improving awareness of cyber threats and contributes to the field through his publications. He can be contacted at: dandy@stikom-bali.ac.id