Jurnal Kecerdasan Buatan dan Pembelajaran Mesin
Vol 1, No 1, 2026, 49-61 https://doi.org/10.35912/jkbpm.v1i1.6764

Air Quality Prediction in DKI Jakarta Using Support Vector Machine: A Comprehensive Classification Approach

Hafidz Tri Utomo Muhammad[1*], Chairani Fauzi[2]
Institut Informatika dan Bisnis Darmajaya, Bandar Lampung, Indonesia[1,2]
hafidz.2111010114@mail.darmajaya.ac.id[1*],

Article History
Received on 13 December 2025
1st Revision on 23 December 2025
2nd Revision on 02 January 2026
3rd Revision on 14 January 2026
Accepted on 29 January 2026

Abstract

Purpose: This study aimed to develop a machine learning-based classification model for predicting Air Pollution Standard Index (ISPU/AQI) categories in DKI Jakarta using the Support Vector Machine (SVM) algorithm. The study addressed the challenge of accurate multiclass air quality classification under conditions of pollution variability and imbalanced class distributions.

Research Methodology: A quantitative research design was employed using 1,825 daily observations obtained from the Satu Data Jakarta portal between February and November 2023. Six pollutant parameters (PM2.5, PM10, CO, SO_2, NO_2, and O_3) were used as predictor variables. Data preprocessing included missing value handling, duplicate removal, label encoding, and min-max normalization. An SVM classifier with a Radial Basis Function (RBF) kernel was implemented and optimized using GridSearchCV with stratified K-fold cross-validation.

Results: The optimized SVM-RBF model achieved an overall classification accuracy of 96.1% on the test dataset. The model showed strong performance for the Good, Moderate, and Unhealthy categories, while classification effectiveness for minority classes was reduced due to severe class imbalance.

Conclusions: The findings demonstrate that SVM-RBF is highly effective for AQI classification and can support the development of automated real-time air quality monitoring systems.

Limitations: The study was limited to DKI Jakarta and a ten-month observation period. The underrepresentation of Very Unhealthy and Hazardous categories constrained model performance for rare pollution events.

Contributions: This study provides a validated SVM-based AQI classification framework and offers methodological guidance for environmental monitoring and urban air quality management in developing countries.

Keywords: *Air Quality Classification, DKI Jakarta, GridSearchCV, Machine Learning, Support Vector Machine*

How to Cite: Muhammad, H. T. U., & Fauzi, C.. (2026). Air Quality Prediction in DKI Jakarta Using Support Vector Machine: A Comprehensive Classification Approach. *Jurnal Kecerdasan Buatan dan Pembelajaran Mesin*, 1(1), 49-61.

1. Introduction

Air quality has emerged as one of the foremost environmental and public health challenges of the twenty first century. The World Health World (2018) estimates that ambient air pollution contributes

to approximately seven million premature deaths annually, with the burden disproportionately borne by populations in low and middle income countries concentrated in densely urbanized areas. In Southeast Asia, rapid industrialization, explosive motorization, and unregulated urban expansion have compounded existing pollution challenges, rendering several regional megacities among the most polluted in the world (IQAir, 2023).

DKI Jakarta, Indonesia capital and primary economic center, epitomizes these challenges. With a population exceeding 11 million within the city boundary and a metropolitan population surpassing 30 million, Jakarta generates substantial pollutant emissions from road transportation, industrial facilities, energy generation, and biomass burning (Putri & Suwanda, 2023). World air quality report consistently ranked Jakarta among the five most polluted capital cities globally, with annual mean PM2.5 concentrations exceeding the WHO guideline levels of 5 $\mu g/m^3$ by more than tenfold (IQAir, 2023). The resultant health burden includes an elevated incidence of acute respiratory infections, chronic obstructive pulmonary disease, cardiovascular disease, and neurodevelopmental effects in children (Alfian, Nurhidayat, & Hidayat, 2024; (World, 2021).

Effective management of urban air pollution requires timely, accurate, and granular information on the air quality status. Indonesia regulatory framework employs the Air Pollution Standard Index (ISPU) a national Air Quality Index (AQI) system administered by the Ministry of Environment and Forestry (Kementerian, 2020) to categorize pollution levels across five tiers: Good, Moderate, Unhealthy, Very Unhealthy, and Hazardous. However, conventional monitoring infrastructure relies on fixed regulatory stations with limited spatial coverage and retrospective reporting, creating temporal gaps that constrain proactive public health communication and policy responses (Rahmawati, Siregar, & Zulkarnain, 2022).

The advent of Machine Learning (ML) methodologies in environmental science has opened up compelling avenues for addressing these limitations. Unlike physics-based atmospheric dispersion models, which demand substantial computational resources and site-specific parametrization, ML algorithms can learn complex nonlinear relationships between pollutant inputs and AQI outcomes directly from empirical data (Zhang, Bocquet, Mallet, Seigneur, & Baklanov, 2012). Among the repertoire of supervised learning algorithms, support vector machines (SVM) have demonstrated consistent superiority in classification tasks involving high-dimensional, non-linearly separable, and relatively limited datasets conditions commonly encountered in environmental monitoring contexts (Chaloulakou, Grivas, & Spyrellis, 2003; Zhu, Wang, Yang, Zhang, Zhang, Ren, Wu, & Ye, 2022).

Despite a growing body of literature applying ML to air quality prediction globally, several research gaps persist in Indonesia. First, most existing domestic studies employ single algorithms without rigorous hyperparameter optimization, limiting the reliability of the reported accuracy metrics (Hasyim, Rahman, & Sutoyo, 2021; Prasetyo, & Nugroho, 2021). Second, few studies apply systematic preprocessing pipelines, including label encoding, feature scaling, and stratified cross-validation, within a unified analytical framework, making it difficult to isolate the contribution of individual methodological components to model performance. Third, the class imbalance problem inherent to AQI data, where extreme pollution categories are structurally rare, has received insufficient attention in the Indonesian literature, despite its practical significance for emergency response planning.

This study addresses these gaps by developing and evaluating an optimized SVM classification model for ISPU prediction in DKI Jakarta, grounded in a comprehensive, reproducible methodology. Specifically, this study: (1) constructs a preprocessing pipeline integrating missing value handling, label encoding, min-max normalization, and stratified train-test splitting; (2) optimizes SVM kernel parameters (C and γ) using GridSearchCV with stratified K-fold cross-validation; (3) evaluates model performance using multi-class precision, recall, F1-score, and confusion matrix metrics; and (4) critically discusses the implications of class imbalance for model reliability in operational deployment (Adawiyah & Iskandar, 2022).

The primary novelty of this research lies in its systematic end-to-end methodological transparency, from raw data acquisition through preprocessing, model training, hyperparameter tuning, and multi-metric evaluation, applied to Jakarta urban air quality context. This reproducibility-oriented approach addresses a recognized weakness in the regional ML-for-environment literature and provides a validated methodological template for researchers and practitioners in Indonesian cities. The remainder of this paper is organized as follows: Section 2 reviews the theoretical and empirical literature; Section 3 details the research methodology; Section 4 presents and discusses the results; and Section 5 concludes with limitations and future research directions (Arias, Bellouin, Coppola, Jones, Krinner, & Marotzke, 2021).

2. Literature Review

2.1 Urban Air Pollution and the AQI Framework

Urban air pollution arises from the interaction of primary emissionsdirectly released from combustion sourcesand secondary pollutants formed through atmospheric chemical reactions (Zhang, Bocquet, Mallet, Seigneur, & Baklanov, 2012; Wen, Liu, Yao, Peng, Li, Hu, & Chi, 2019). The six key pollutants monitored by Indonesia ISPU system PM2.5, PM10, CO, SO_2, NO_2, and O_3 represent a cross-section of combustion by-products and photochemical oxidants, each with distinct health implications. PM2.5 and PM10 are particularly critical because of their deep respiratory penetration and association with cardiovascular and pulmonary mortality (World, 2021; Tsai, Zeng, & Chang, 2018). AQI frameworks, such as the ISPU, aggregate multipollutant data into a single communicable index, enabling public health advisories calibrated to pollution severity (Kementerian, 2020).

2.2 Machine Learning for Air Quality Classification

The application of ML to air quality prediction has evolved substantially since early comparative studies that pitted regression models against neural networks for PM10 forecasting (Chaloulakou, Grivas, & Spyrellis, 2003). Subsequent advances have explored ensemble methods, deep learning architectures, and hybrid approaches that integrate atmospheric physics with data-driven components (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, & Duchesnay, 2011) . A comprehensive review by Zhang, Bocquet, Mallet, Seigneur, and Baklanov (2012) documented the rapid evolution of real-time air quality forecasting frameworks, emphasizing the role of data quality, feature selection, and model validation protocols in determining predictive reliability (World, 2021).

Within the classification paradigm, SVM has received sustained attention in air quality applications. Its theoretical foundation in statistical learning theory particularly the structural risk minimization principle confers robustness advantages over empirical risk minimization algorithms, especially under limited training data conditions (Chen, & Guestrin, 2016; Hastie, Tibshirani, & Friedman, 2009). The kernel trick, which implicitly maps inputs into high dimensional feature spaces, enables the SVM to construct nonlinear decision boundaries without explicit feature engineering, which is a critical capability given the complex multivariate interactions among atmospheric pollutants (Sari, & Wibowo, 2023; Suryani, Wulandari, & Putri, 2021; Nguyen, Cai, & Bhatt, 2023).

2.3 SVM Applications in Environmental and Air Quality Research

Internationally, SVM has been applied to diverse air quality tasks, including NO_2 concentration forecasting, PM2.5 prediction, and multi-class AQI classification (Zhu, Wang, Yang, Zhang, Zhang, Ren, Wu, & Ye, 2022). Compared to competing algorithms such as Random Forest and Gradient Boosting, SVM demonstrates competitive accuracy with superior interpretability in binary and multi-class classification tasks on structured environmental datasets (Li, Peng, Yao, Cui, Hu, You, & Chi, 2017).

In the Indonesian context, Hasyim, Rahman, and Sutoyo (2021) conducted a direct comparison of SVM and Naïve Bayes classifiers for air quality categorization, demonstrating SVM superiority with an accuracy of 89.3% versus 76.4% for Naïve Bayes. Prasetyo and Nugroho (2021) applied SVM alongside Decision Tree classifiers for urban air quality prediction, reporting that SVM performance was highly sensitive to kernel selection and regularization parameter C, underscoring the importance

of systematic hyperparameter optimization. Suryani, Wulandari, and Putri (2021) demonstrated that ML approaches, including SVM, outperformed statistical time-series models for categorical AQI prediction. Utami and Nurfikri (2021) applied SVM for air quality data classification, emphasising the importance of multi-metric evaluation beyond accuracy alone (Kumar, Gulia, Harrison, & Khare, 2017).

2.4 Identified Research Gaps

A synthesis of the literature reveals three principal gaps that motivate the present study. First, existing Indonesian SVM-based air quality studies predominantly report accuracy as a solitary evaluation metric, obscuring differential model performance across AQI categories, which is a critical omission given the class-imbalanced nature of AQI datasets (Utami & Nurfikri, 2021). Second, preprocessing pipelines are frequently underspecified in published studies, preventing reproducibility assessments and methodological comparisons across studies. Third, the specific context of Jakarta ISPU dataset from the Satu Data Jakarta portal remains underexplored using rigorous ML methodology, despite its public availability and policy relevance. This study directly addresses each of these gaps.

2.5 Theoretical Framework

The theoretical framework of this study integrates two complementary perspectives. From the computational intelligence perspective, SVM is grounded in Vapnik Chervonenkis (VC) theory, which formalizes the relationship between model complexity, training error, and generalization performance (Hastie, Tibshirani, & Friedman, 2009). The bias-variance trade-off, operationalized through regularization parameter C and kernel bandwidth γ, is central to this framework. From the environmental informatics perspective, this study conceptualizes air quality classification as a supervised pattern recognition problem in which atmospheric state vectors (pollutant concentrations) are mapped to regulatory risk categories, enabling automated decision support for environmental management agencies (Rahmawati, Siregar, & Zulkarnain, 2022; Zhang, Bocquet, Mallet, Seigneur, & Baklanov, 2012).

3. Research Methodology

3.1 Research Design

This study adopted a quantitative, design-and-evaluate research methodology structured around the development and validation of a supervised machine learning classification model. The methodology followed a sequential pipeline: data acquisition, preprocessing, feature engineering, model development, hyperparameter optimization, evaluation, and interpretation.

3.2 Data Source and Description

Secondary data were obtained from the Satu Data Jakarta open government data portal (satudata.jakarta.go.id), specifically the ISPU dataset for the DKI Jakarta Province. The dataset encompasses daily observations from five monitoring stations across the city, covering the period from February 2023 to November 2023. After merging and deduplication, the final dataset comprised 1,825 records. Each record contains measurements of six pollutant parameters (PM2.5, PM10, CO, SO_2, NO_2, and O_3) alongside an ISPU categorical label assigned according to KLHK's (2020) classification thresholds. All data processing and modelling were conducted on the Google Colaboratory cloud platform using Python 3 with the Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn libraries (Lundberg & Lee, 2017).

3.3 Data Preprocessing

The preprocessing pipeline comprised four sequential steps. First, missing value handling involved the identification and removal of records with null values across any feature column, as the proportion of missing data was sufficiently low (< 2%) to justify listwise deletion without introducing significant bias. Second, duplicate detection was performed to eliminate repeated observations arising from data export artifacts. Third, label encoding converted the categorical ISPU target variable into integer representations: good = 0, moderate = 2, unhealthy = 3, Very Unhealthy = 4, and hazardous = 5. Fourth, Min-Max normalization was applied to all six continuous predictor features to rescale values to the [0, 1] interval using the following formula:

$$X' = (X - X_{min}) / (X_{max} - X_{min})$$

This normalization step is essential to prevent features with larger absolute ranges (e.g., CO measured in μg/m³) from disproportionately influencing SVM kernel distance calculations Utami & Nurfikri, 2021. The dataset was subsequently partitioned into training (80%, n = 1,460) and test (20%, n = 365) subsets using stratified random sampling to preserve the proportional class distribution across both the splits.

3.4 Model Evaluation

The model performance was evaluated on the held-out test set (n = 365) using the following metrics:

1. Overall Accuracy: Proportion of correctly classified instances across all classes.
2. Per-class Precision: Positive predictive value for each ISPU category.
3. Per-class Recall sensitivityy): True positive rate for each ISPU category.
4. Per-class F1-Score: Harmonic mean of precision and recall, providing a balanced performance indicator.
5. Macro-e and weighted-average metrics:These provide unweighted and class-frequency-weighted aggregations, respectively.
6. Confusion matrix: Visualization of the full distribution of predicted versus actual class assignments.

The distinction between macro-average and weighted-average metrics is particularly informative in the presence of class imbalance: the macro-average treats all classes equally, whereas the weighted-average accounts for class frequency, potentially masking poor minority class performance (Hastie, Tibshirani, & Friedman, 2009).

3.5 Model Development: Support Vector Machine

The SVM classifier was implemented using the SVC class of Scikit learn. The Radial Basis Function (RBF) kernel was selected based on its theoretical suitability for non-linearly separable multi-class classification problems, consistent with prior applications in environmental data contexts (Hasyim, Rahman, & Sutoyo, 2021; Prasetyo, & Nugroho, 2021). The multi-class extension employed a One Versus One (OvO) strategy, which constructs $k(k-1)/2$ binary classifiers for k classes and determines the final class assignment by majority voting.

The selection of the RBF kernel is particularly advantageous for air quality classification because pollutant interactions often exhibit complex and non-linear relationships. Atmospheric pollution levels are influenced by multiple factors, including meteorological conditions, traffic density, industrial activities, and photochemical reactions, which rarely follow simple linear patterns. By mapping the input features into a higher-dimensional space, the RBF kernel enables the SVM model to capture these intricate relationships more effectively than linear classifiers. As a result, the model can establish more flexible decision boundaries that improve its ability to distinguish between AQI categories with overlapping pollutant concentration ranges.

Moreover, the One Versus One (OvO) multiclass strategy provides several practical benefits for AQI classification tasks involving multiple pollution categories. Because each binary classifier focuses on distinguishing between only two classes at a time, the resulting decision boundaries are often more precise and computationally efficient than those produced by direct multiclass approaches. This characteristic is particularly valuable when class distributions vary substantially, as observed in the present dataset. The majority-voting mechanism further enhances classification robustness by aggregating predictions from multiple classifiers, thereby reducing the influence of individual classification errors and contributing to the high overall predictive performance achieved by the model.

3.6 Hyperparameter Optimisation

Systematic hyperparameter optimization was conducted using Scikit-learn GridSearchCV procedure. The parameter search space included the regularization parameter $C \in \{0.1, 1, 10, 100\}$ and kernel coefficient $\gamma \in \{0.001, 0.01, 0.1, 1\}$. The search evaluated 16 parameter combinations, each assessed using 5-fold stratified cross-validation on the training set to ensure a proportional class representation within each fold. The combination that yielded the highest mean cross-validation accuracy was selected as the final model configuration. The random state was fixed at 42 to ensure the reproducibility of the train-test split and fold assignments.

4. Results and Discussions

4.1 Dataset Characteristics and Class Distribution

The 1,825-record ISPU dataset exhibited a notably imbalanced class distribution. The Moderate category accounted for the substantial majority of observations (approximately 74.3% of the total dataset), reflecting Jakarta chronic moderate pollution baseline. The Good category constituted approximately 12.8% of the records, whereas the Unhealthy category accounted for approximately 12.4%. The Very Unhealthy and Hazardous categories collectively represent fewer than 1% of observations, corresponding to episodic extreme pollution events. This structural imbalance carries important implications for model evaluation, as a trivial classifier predicting 'Moderate' for all instances would achieve an accuracy exceeding 74% without any genuine predictive capability.

The six pollutant features exhibited varying distributional properties. PM2.5 and PM10 displayed right-skewed distributions that were consistent with episodic pollution spikes. CO concentrations showed a strong correlation with PM2.5 ($r \approx 0.61$), consistent with the shared combustion emission sources. O_3 concentrations exhibited a distinct diurnal photochemical formation pattern, as evidenced by higher afternoon values. Following Min-Max normalization, all features were successfully rescaled to [0, 1].

4.2 Hyperparameter Optimisation Results

GridSearchCV identified the optimal hyperparameter combination as $C = 10$ and $\gamma = 0.1$ for the RBF kernel, yielding a mean cross-validation accuracy of 95.3% (SD = 0.8%) across the five training folds. This finding is consistent with the general guidance that moderate regularization ($C = 10$) balanced with an intermediate kernel bandwidth ($\gamma = 0.1$) prevents both underfitting and overfitting in multi-class SVM applications with normalized features (Hastie, Tibshirani, & Friedman, 2009). Table 1 summarizes the cross-validation performance of the top-performing hyperparameter combinations.

Table 1. Cross-Validation Performance Across Selected Hyperparameter Combinations (* = Optimal)

C	γ (Gamma)	Kernel	CV Accuracy (%)	SD (%)
0.1	0.1	RBF	87.4	1.2
1	0.1	RBF	92.8	1.0
10	0.01	RBF	93.6	0.9
10	0.1	RBF	95.3*	0.8
100	0.1	RBF	94.9	1.1
100	1	RBF	93.1	1.3

Table 1 presents the results of hyperparameter tuning for the Support Vector Machine (SVM) model using the Radial Basis Function (RBF) kernel with different combinations of the penalty parameter (C) and gamma (γ). The findings indicate that model performance is highly influenced by the selection of these parameters. The lowest cross-validation accuracy (87.4%) was obtained when both C and γ were set to 0.1, suggesting insufficient model complexity for capturing the underlying data patterns. Increasing the value of C generally improved classification performance, with accuracy rising to 92.8% for $C = 1$ and $\gamma = 0.1$, and further to 93.6% for $C = 10$ and $\gamma = 0.01$. The highest cross-validation accuracy of 95.3% was achieved with $C = 10$ and $\gamma = 0.1$, accompanied by the lowest standard deviation (0.8%), indicating both strong predictive performance and stable results across

validation folds. However, further increasing C to 100 did not yield additional improvements, as accuracy slightly decreased to 94.9% and 93.1% for γ values of 0.1 and 1, respectively. These results suggest that the optimal parameter combination for this dataset is C = 10 and γ = 0.1, providing the best balance between model generalization and classification accuracy.

4.3 Classification Performance on the Test Set

The optimized SVM model achieved an overall test accuracy of 96.1%, substantially exceeding the naïve majority-class baseline of 74.3% and confirming its genuine discriminative capability. Table 2 presents the per-class and aggregate performance metrics of the models.

Table 2. Per-Class Classification Performance Metrics on the Test Set (n = 365)

AQI Category	Precision (%)	Recall (%)	F1-Score (%)	Support (n)
Good (0)	84	88	86	47
Moderate (2)	98	98	98	271
Unhealthy (3)	95	91	93	45
Very Unhealthy (4)	100	50	67	2
Hazardous (5)	—	—	—	0
Macro Average	94	82	86	365
Weighted Average	96	96	96	365

Table 2 shows the Moderate category achieved the highest performance (F1 = 98%), reflecting its dominant representation in the training set. The Unhealthy category also demonstrated a robust performance (F1 = 93%), indicating that the model captured the characteristic pollutant concentration patterns differentiating unhealthy from moderate conditions. The Good category showed a slightly lower recall (88%) relative to precision (84%), suggesting the occasional misclassification of genuinely clean-air days into the Moderate category, likely attributable to overlapping PM2.5 concentration ranges at the Good-Moderate boundary.

The Very Unhealthy category exhibited a pronounced precision-recall asymmetry (precision = 100%, recall = 50%), indicating that while the model's positive predictions for this category were perfectly reliable, it missed half of the actual Very Unhealthy instances. This pattern is a classical consequence of severe class imbalance in the training data (n = 2 instances in the test set) and represents the most significant operational limitation of the current model. The hazardous category was entirely absent from the test split, precluding its evaluation.

The gap between the macro-average (F1 = 86%) and weighted-average (F1 = 96%) metrics quantitatively captures the class imbalance effect: the macro-average, which weights all categories equally, is substantially lower than the weighted-average, which is dominated by the Moderate category. This divergence underscores the importance of reporting both metrics and not relying on accuracy or weighted-average F1 alone as indicators of model quality for AQI classification.

4.4 Comparative Analysis with Prior Studies

The 96.1% overall accuracy achieved in this study compares favorably with previous Indonesian SVM-based air quality classification studies. Hasyim, Rahman, and Sutoyo (2021) reported 89.3% accuracy on a smaller dataset without systematic hyperparameter optimisation, while Prasetyo and Nugroho (2021) achieved 91.4% with SVM on a different regional dataset. The performance improvement in the present study is attributable to (1) the application of grid search CV for systematic hyperparameter optimization, (2) the use of stratified cross-validation to ensure representative fold composition, and (3) the comprehensive preprocessing pipeline minimizing noise-induced model interference. These findings reinforce the conclusion of Suryani, Wulandari, and Putri (2021) that methodological rigor is a primary determinant of ML model performance, beyond algorithm selection alone.

Furthermore, the findings highlight the importance of model development procedures in achieving high predictive performance for environmental datasets. While algorithm selection remains an important consideration, the results demonstrate that preprocessing quality, feature normalization, and systematic parameter tuning can substantially influence classification outcomes. The strong performance achieved by the SVM-RBF model suggests that carefully optimized conventional machine learning algorithms may remain competitive with more complex approaches, particularly when datasets are of moderate size and feature relationships are non-linear. This observation is consistent with recent studies indicating that rigorous data preparation and validation strategies often contribute more to predictive success than the adoption of increasingly sophisticated algorithms.

The results also underscore the value of reproducible and transparent machine learning workflows in environmental informatics research. By employing established procedures such as GridSearchCV and stratified cross-validation, the study reduces the risk of biased performance estimation and enhances the reliability of the reported results. Such methodological transparency is particularly important for applications involving public health and environmental policy, where decision-makers require confidence in model outputs before integrating them into operational systems. Consequently, future studies should prioritize standardized evaluation frameworks and comparative benchmarking to ensure that reported performance gains reflect genuine improvements in predictive capability rather than differences in experimental design or dataset composition.

4.5 Implications for Real-Time Air Quality Monitoring

The demonstrated accuracy and computational efficiency of the SVM-RBF model support its integration into operational air-quality monitoring architectures. The model's inference speed, typically under 100 ms for batch prediction with Scikit-learn SVC implementation, is compatible with real-time data processing pipelines fed by IoT sensor networks (Rahmawati, Siregar, & Zulkarnain, 2022). Potential deployment pathways include integration with Jakarta existing regulatory station network for automated ISPU nowcasting, mobile application platforms providing location-specific pollution advisories, and early warning systems that trigger public health communications when model predictions indicate imminent category escalation. However, operational deployment must be conditioned on the resolution of the class imbalance limitation described above, particularly given the life-safety implications of failing to detect Very Unhealthy or Hazardous conditions.

In addition, the successful implementation of the SVM-RBF model has important implications for evidence-based environmental governance and urban sustainability planning. Accurate and timely air quality classification can support government agencies in identifying pollution hotspots, evaluating the effectiveness of emission control policies, and allocating resources more efficiently. By transforming large volumes of environmental monitoring data into actionable information, machine learning-based systems can enhance decision-making processes and contribute to the achievement of public health and environmental protection objectives. Furthermore, the availability of near real-time AQI predictions may encourage greater public awareness and behavioral adaptation, such as reducing outdoor activities during periods of elevated pollution or adopting preventive health measures for vulnerable populations.

From a broader technological perspective, the integration of machine learning models into smart city infrastructures represents a significant step toward data-driven environmental management. As urban sensor networks continue to expand, the volume and granularity of air quality data will increase substantially, creating opportunities for more sophisticated predictive analytics and adaptive monitoring systems. The proposed SVM framework may serve as a foundational component within a larger ecosystem that combines meteorological forecasting, traffic monitoring, and industrial emission tracking to provide comprehensive environmental intelligence. Such integration could facilitate proactive interventions, improve urban resilience to pollution-related risks, and support long-term strategies aimed at enhancing air quality and quality of life in rapidly growing metropolitan regions.

4.6 Discussions

The results demonstrate that the Support Vector Machine (SVM) with an RBF kernel is highly effective for classifying Jakarta's Air Quality Index (AQI) categories despite the substantial class imbalance present in the dataset. The dominance of the Moderate category reflects the typical air quality conditions experienced in Jakarta and creates a challenging classification environment where a naïve classifier could achieve relatively high accuracy without meaningful predictive power. Nevertheless, the proposed model achieved a test accuracy of 96.1%, substantially outperforming the majority-class baseline of 74.3%, indicating that the model successfully learned discriminative patterns among pollutant concentrations rather than relying on class frequency alone. The observed correlations among pollutants, particularly between PM2.5 and CO, further support the suitability of machine learning approaches for capturing complex environmental relationships that may not be evident through traditional threshold-based classification methods (Sarkar, Singh, & Kumar, 2023).

The hyperparameter optimization process played a critical role in achieving the reported performance. GridSearchCV identified $C = 10$ and $\gamma = 0.1$ as the optimal parameter combination, producing the highest cross-validation accuracy and the lowest variability across folds. This finding highlights the importance of balancing model complexity and generalization capability when applying SVMs to environmental datasets. Lower values of C resulted in underfitting and reduced predictive performance, whereas excessively large values increased the risk of overfitting without providing meaningful gains in accuracy. The relatively small standard deviation observed across validation folds suggests that the optimized model maintained stable performance and was not overly sensitive to variations in the training data, reinforcing the reliability of the selected configuration (Ketu, & Mishra, 2021; Saminathan, & Malathy, 2023).

The class-specific evaluation metrics provide additional insight into the strengths and weaknesses of the model. Excellent performance was achieved for the Moderate and Unhealthy categories, demonstrating the model's ability to distinguish between common pollution conditions with high precision and recall. However, the substantially lower recall for the Very Unhealthy category reveals a persistent challenge associated with severe class imbalance. Although all predictions assigned to this category were correct, the model failed to identify half of the actual Very Unhealthy instances (Rosales-Pérez, García, & Herrera, 2022). This limitation is particularly important in the context of public health monitoring, where missed detections of extreme pollution events may delay warning dissemination and risk mitigation efforts. The discrepancy between macro-average and weighted-average performance metrics further emphasizes the necessity of evaluating models using multiple criteria rather than relying solely on overall accuracy (Méndez, Merayo, & Núñez, 2023).

Compared with previous Indonesian air quality classification studies, the proposed approach achieved superior predictive performance, supporting the view that methodological rigor is a key determinant of machine learning success. The combination of systematic hyperparameter optimization, stratified cross-validation, and comprehensive preprocessing contributed to the observed improvement over earlier SVM-based implementations. These findings suggest that SVM-RBF models can serve as a reliable foundation for real-time air quality monitoring systems and automated AQI prediction platforms. However, before operational deployment, future research should address the imbalance problem through techniques such as Synthetic Minority Over-sampling Technique (SMOTE), cost-sensitive learning, or ensemble-based approaches to improve the detection of rare but critical pollution categories. Such enhancements would increase the robustness of the system and strengthen its practical value for environmental management and public health protection (Chandra, Suprihatin, & Resti, 2023).

5. Conclusions

5.1 Conclusion

This study successfully developed and validated a Support Vector Machine classification model for predicting Air Pollution Standard Index (ISPU) categories in DKI Jakarta using six-pollutant daily monitoring data from the Satu Data Jakarta portal. Through a systematic methodology encompassing preprocessing, min-max normalization, stratified train-test splitting, RBF kernel SVM training, and

GridSearchCV hyperparameter optimization, the model achieved an overall classification accuracy of 96.1% on the held-out test set.

The per-class analysis revealed strong performance for the dominant Moderate (F1 = 98%) and Unhealthy (F1 = 93%) categories, with moderate performance for the Good category (F1 = 86%). The primary limitation was the poor recall for the Very Unhealthy class (50%), directly attributable to the structural class imbalance in the training data. The discrepancy between the macro-average (F1 = 86%) and weighted-average (F1 = 96%) metrics highlights the hazard of relying solely on overall accuracy for imbalanced multiclass classification evaluation.

The methodological contribution of this research lies in its reproducible, end-to-end pipelinefrom raw governmental open data through multi-metric model evaluationapplied to a policy-relevant urban air quality context. The findings establish a validated benchmark for SVM-based AQI classification in Indonesian metropolitan settings and provide a methodological template for similar studies across the country.

5.2 Research Limitations

Several limitations constrain the conclusions that can be drawn from this study. First, the dataset is geographically confined to DKI Jakarta and temporally limited to a ten-month window (February–November 2023), potentially missing seasonal pollution dynamics characteristic of the dry-wet season transition typical of Java tropical climate. Second, the severe class imbalance in extreme AQI categories (Very Unhealthy, Hazardous) fundamentally limits model reliability for the highest-risk pollution events precisely those events for which accurate prediction carries the greatest public health significance. Third, the study did not implement data balancing techniques, such as SMOTE, which may have improved minority class performance. Fourth, no comparison was made with alternative classification algorithms (e.g., Random Forest, XGBoost, LSTM), precluding an assessment of whether SVM represents the optimal algorithmic choice for this specific dataset. Fifth, external validation using independent datasets from other Indonesian cities was not conducted.

Another limitation relates to the exclusive reliance on pollutant concentration variables as model inputs. Although PM2.5, PM10, CO, SO_2, NO_2, and O_3 are the primary determinants of AQI classification, air quality is also influenced by external factors such as temperature, humidity, wind speed, rainfall intensity, and traffic volume. The exclusion of these meteorological and anthropogenic variables may have reduced the model's ability to capture complex environmental interactions that contribute to fluctuations in pollution levels. Consequently, the predictive performance reported in this study may not fully reflect the potential accuracy achievable through a more comprehensive feature set incorporating both pollutant and contextual environmental data.

In addition, the evaluation focused primarily on classification performance metrics derived from a single train-test split and cross-validation procedure. While these methods provide a reliable assessment of model effectiveness, they do not fully capture the long-term operational challenges associated with real-world deployment. Factors such as sensor measurement errors, missing data streams, changes in emission patterns, and evolving urban environmental conditions may affect model stability over time. Therefore, future studies should investigate model robustness under dynamic operational scenarios and conduct longitudinal validation to ensure that predictive performance remains consistent when applied to continuously updated air quality monitoring systems.

5.3 Suggestions and Directions for Future Research

Future research should focus on improving model performance, particularly for rare but critical pollution categories. The severe class imbalance observed in the Very Unhealthy and Hazardous categories can be addressed through advanced data balancing techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) or class-weighted learning approaches, thereby enhancing the model's ability to detect extreme pollution events with significant public health implications. In addition, comparative studies involving alternative machine learning algorithms, including ensemble methods such as Random Forest, XGBoost, and LightGBM, as well as deep

learning architectures such as Long Short-Term Memory (LSTM) networks and Temporal Convolutional Networks (TCNs), would provide valuable insights into the relative strengths and weaknesses of different approaches for urban AQI classification. Expanding the dataset to include multiple years and complete seasonal cycles would also enable a more comprehensive assessment of temporal stability, particularly during dry-season biomass burning episodes that periodically contribute to severe air pollution in the Jakarta metropolitan area.

Furthermore, future studies should explore broader implementation and operational enhancements of the proposed framework. Applying the methodology to other major Indonesian cities, such as Surabaya, Bandung, and Medan, would help evaluate its generalizability and support the development of a standardized national AQI classification system. The integration of low-cost Internet of Things (IoT) sensor networks, edge computing technologies, and real-time inference platforms could facilitate the deployment of an automated air quality monitoring and early warning system capable of providing timely pollution alerts. Additionally, incorporating explainable artificial intelligence techniques, such as SHAP (SHapley Additive exPlanations), would improve model transparency by identifying the pollutant variables that most strongly influence AQI category transitions. Such insights could support evidence-based environmental policies and more targeted emission reduction strategies.

Acknowledgement
The authors gratefully acknowledge the Environmental Agency (Dinas Lingkungan Hidup) of DKI Jakarta Province for providing ISPU monitoring data publicly accessible through the Satu Data Jakarta portal. The authors also extend their appreciation to the Institut Informatika dan Bisnis Darmajaya for the institutional support throughout this research. This study did not receive any external funding.

Author Contributions
HTUM Conceptualization, data collection and curation, preprocessing pipeline development, model training and optimization, results analysis, original draft preparation, and manuscript revision. CF Supervision, conceptualization, methodology review, critical review of results and discussion, and final approval of the submitted version.

References
Adawiyah, R., & Iskandar Mulyana, D. (2022). Optimasi deteksi penyakit kulit menggunakan metode Support Vector Machine (SVM) dan Gray Level Co-occurrence Matrix (GLCM). *INFORMASI (Jurnal Informatika dan Sistem Informasi)*, *14*(1), 18-33. https://doi.org/10.37424/informasi.v14i1.138

Alfian, F., Nurhidayat, M., & Hidayat, T. (2024). Analisis dampak pencemaran udara terhadap kesehatan di wilayah perkotaan. *Jurnal Kesehatan Lingkungan*, *16*(1), 45-52. https://doi.org/10.59966/semar.v2i3.885

Arias, P. A., Bellouin, N., Coppola, E., Jones, R. G., Krinner, G., Marotzke, J., et al. (2021). Technical summary. In V. Masson-Delmotte et al. (Eds.), *In V. Masson-Delmotte et al. (Eds.), Climate Change 2021: The Physical Science Basis (pp. 33–144). Cambridge University Press* (pp. 35-144). Cambridge University Press.

Chaloulakou, A., Grivas, G., & Spyrellis, N. (2003). Neural network and multiple regression models for PM10 prediction in Athens: A comparative assessment. *Journal of the Air & Waste Management Association*, *53*(10), 1183-1190. https://doi.org/10.1080/10473289.2003.10466276

Chandra, W., Suprihatin, B., & Resti, Y. (2023). Median-KNN Regressor-SMOTE-Tomek Links for handling missing and imbalanced data in air quality prediction. *Symmetry*, *15*(4), 887. https://doi.org/10.3390/sym15040887

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Paper presented at In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction (2nd ed.)*.

Hasyim, S. H., Rahman, A., & Sutoyo, T. (2021). Klasifikasi kualitas udara menggunakan metode SVM dan Naïve Bayes. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *5*(3), 511-518. https://doi.org/10.29207/resti.v5i3.3340

IQAir. (2023). 2023 world air quality report: Region and city PM2.5 ranking. *IQAir AG*.

Kementerian Lingkungan Hidup dan Kehutanan (KLHK). (2020). *Peraturan Menteri Lingkungan Hidup dan Kehutanan No*.

Ketu, S., & Mishra, P. K. (2021). Scalable kernel-based SVM classification algorithm on imbalance air quality data for proficient healthcare. *Complex & Intelligent Systems*, *7*(5), 2597-2615. https://doi.org/10.1007/s40747-021-00435-5

Kumar, P., Gulia, S., Harrison, R. M., & Khare, M. (2017). The influence of odd-even car trial on fine and coarse particles in Delhi. *Environmental Pollution, 225, 20–29*. https://doi.org/10.1016/j.envpol.2016.12.037

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., & Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental Pollution, 231, 997–1004*. https://doi.org/10.1016/j.envpol.2017.08.114

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30, 4765–4774*. https://doi.org/10.48550/arXiv.1705.07874

Méndez, M., Merayo, M. G., & Núñez, M. (2023). Machine learning algorithms to forecast air quality: A survey. *Artificial Intelligence Review*, *56*(9), 10031-10066. https://doi.org/10.1007/s10462-023-10424-4

Nguyen, H. T., Cai, W., & Bhatt, D. L. (2023). Machine learning applications in environmental health: A systematic review. *Environmental Research, 220, 115213*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.48550/arXiv.1201.0490

Prasetyo, D., & Nugroho, R. A. (2021). Prediksi kualitas udara menggunakan SVM dan decision tree. *Jurnal Teknik ITS*, *10*(2). https://doi.org/10.12962/j23373539.v10i2.71234

Putri, L. A., & Suwanda. (2023). Implementasi metode Artificial Neural Network (ANN) algoritma backpropagation untuk klasifikasi kualitas udara di Provinsi DKI Jakarta tahun 2021. *Bandung Conference Series: Statistics*, *3*(2), 345-352. https://doi.org/10.29313/bcss.v3i2.7826

Rahmawati, L., Siregar, M., & Zulkarnain, H. (2022). Sistem pemantauan kualitas udara berbasis data dan IoT. *Jurnal Teknologi Informasi dan Komputer*, *10*(3), 210-218..

Rosales-Pérez, A., García, S., & Herrera, F. (2022). Handling imbalanced classification problems with support vector machines via evolutionary bilevel optimization. *IEEE Transactions on Evolutionary Computation. Advance online publication*. https://doi.org/10.48550/arXiv.2204.10231

Saminathan, S., & Malathy, C. (2023). Ensemble-based classification approach for PM2.5 concentration forecasting using meteorological data. *Frontiers in Big Data, 6, Article 1175259*. https://doi.org/10.3389/fdata.2023.1175259

Sari, R. P., & Wibowo, A. (2023). Comparative study of machine learning algorithms for air quality index prediction in Indonesian cities. *Journal of Environmental Informatics*, *41*(1), 55-67. https://doi.org/10.3808/jei.202300477

Sarkar, D., Singh, P., & Kumar, R. (2023). Air quality index prediction using machine learning for Ahmedabad city. *Digital Chemical Engineering, 7, 100093*. https://doi.org/10.1016/j.dche.2023.100093

Suryani, E., Wulandari, A., & Putri, R. D. (2021). Prediksi kualitas udara menggunakan algoritma machine learning. *Jurnal Informatika*, *15*(2), 155-162..

Tsai, Y. T., Zeng, Y.-R., & Chang, Y.-S. (2018). Air pollution forecasting using RNN with LSTM. Paper presented at In 2018 IEEE 16th International Conference on Dependable, Autonomic and Secure Computing (pp. 1074–1079). IEEE.

Utami, W. S., & Nurfikri, F. (2021). Penerapan Support Vector Machine (SVM) untuk klasifikasi data kualitas udara. *JPIT*, *6*(2), 88-94. https://doi.org/10.30591/jpit.v6i2.2399

Wen, C., Liu, S., Yao, X., Peng, L., Li, X., Hu, Y., & Chi, T. (2019). A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. *Science of the Total Environment, 654, 1091–1099*.

World Health Organization. (2018). *Air pollution and child health: Prescribing clean air*.

World Health Organization. (2021). *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., & Baklanov, A. (2012). Real-time air quality forecasting, part I: History, techniques, and current status. *Atmospheric Environment, 60, 632–655*. https://doi.org/10.1016/j.atmosenv.2012.06.031

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environment and Health*, *1*(2), 107-116. https://doi.org/10.1016/j.eehl.2022.06.001