ABNUS JOURNAL
Your Research Partner

Jurnal Teknika dan Informatika (JTI)
E-ISSN: 2797-0000 | Vol. 6, No. 1 (2026) | pp. 12-26

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

Hafidz Tri Utomo Muhammad[1*], Chairani Fauzi[2]

Darmajaya Institute of Informatics and Business, Bandar Lampung, Indonesia[1,2]

***Corresponding author:** hafidz.2111010114@mail.darmajaya.ac.id |*

Received: 9 March 2026 | Revised: 24 April 2026 | Published: 25 May 2026

Abstract

Purpose: This study develops a machine learning-based classification model to predict Air Pollution Standard Index (ISPU/AQI) categories in DKI Jakarta using the Support Vector Machine (SVM) algorithm. It addresses the challenge of accurate multi-class air quality classification under highly variable urban pollution conditions and imbalanced class distributions.
Research Methodology: A quantitative approach was employed using 1,825 daily observations from the Satu Data Jakarta portal collected between February and November 2023. Six pollutant variables (PM2.5, PM10, CO, SO_2, NO_2, and O_3) were used as predictors. Data preprocessing included missing value imputation, duplicate removal, label encoding, and Min-Max normalization. An SVM model with a Radial Basis Function (RBF) kernel was implemented in Python (Scikit-learn) on Google Colaboratory. Hyperparameters were optimized using GridSearchCV with StratifiedKFold cross-validation. Model performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrix analysis.
Results: The optimized SVM-RBF model achieved 96.1% classification accuracy on the test set. Macro-average F1-scores reached 86% for the Good category, 98% for Moderate, and 93% for Unhealthy. Performance for Very Unhealthy and Hazardous categories was lower due to class imbalance and limited observations.
Conclusions: The proposed model provides accurate and reliable AQI classification for urban tropical environments and offers a practical foundation for automated real-time air quality monitoring systems.
Limitations: The study is limited to DKI Jakarta and a ten-month observation period, while minority classes remain underrepresented.
Contributions: This study presents a validated SVM-based AQI classification framework that can support scalable, data-driven air quality monitoring and environmental decision-making in Indonesian metropolitan areas.

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 Teknika dan Informatika (JTI)*, 6(1), 12-26. https://doi.org/10.52909/jti.v6i1.324

1. Introduction

Air quality has emerged as one of the foremost environmental and public health challenges of the twenty-first century, particularly in rapidly urbanizing regions where industrialization, transportation, and population growth have substantially increased atmospheric pollutant emissions. Poor air quality is associated with a wide range of adverse health outcomes, including respiratory and cardiovascular diseases, reduced life expectancy, and increased premature mortality. The World Health Organization 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 urbanised areas (World, 2018). In Southeast Asia, rapid industrialisation, explosive motorisation, 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 centre, epitomises these challenges. With a population exceeding eleven million within the city boundary and a metropolitan population surpassing thirty million, Jakarta generates substantial pollutant emissions from road transportation, industrial facilities, energy generation, and biomass burning. These emission sources, combined with rapid urban expansion, dense traffic congestion, and changing meteorological conditions, contribute to considerable spatial and temporal variations in ambient air quality. Episodes of elevated pollutant concentrations occur frequently, particularly during periods of limited atmospheric dispersion, increasing the risk of exposure for millions of residents. IQAir 2023 World Air Quality Report consistently ranked Jakarta among the five most polluted capital cities globally, with annual mean PM 2.5 concentrations exceeding WHO guideline levels of 5 µg/m³ by more than tenfold. The resultant health burden includes elevated incidence of acute respiratory infections, chronic obstructive pulmonary disease, cardiovascular disease, and neurodevelopmental effects in children (World, 2018).

Effective management of urban air pollution requires timely, accurate, and granular information on air quality status. Indonesia regulatory framework employs the *Indeks Standar Pencemar Udara* (ISPU) —a national Air Quality Index (AQI) system administered by the Ministry of Environment and Forestry (Kementerian, 2020)—to categorise pollution levels across five tiers: Good, Moderate, Unhealthy, Very Unhealthy, and Hazardous. These categories provide a standardized interpretation of ambient air quality by translating pollutant concentrations into health-related risk levels that are easily understood by policymakers and the general public. The ISPU serves not only as a communication tool for issuing health advisories but also as an important reference for environmental monitoring, regulatory compliance, and pollution mitigation planning. As air quality conditions may change rapidly due to fluctuations in emissions and meteorological factors, the ability to classify AQI categories accurately and efficiently is essential for supporting early warning systems and evidence-based environmental decision-making.. 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 response (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

The advent of machine learning (ML) methodologies in environmental science has opened compelling avenues for addressing these limitations (Du, Bai, Tan, Xue, Samat, Xia, & Liu, 2020). Unlike physics-based atmospheric dispersion models, which demand substantial computational resources and site-specific parameterisation, ML algorithms can learn complex non-linear relationships between pollutant inputs and AQI outcomes directly from empirical data (Zhang, Bocquet, Mallet, Seigneur, & Baklanov, 2012). This capability enables the development of predictive models that are computationally efficient, highly scalable, and adaptable to diverse environmental conditions without requiring explicit knowledge of atmospheric processes. As the availability of high-frequency air quality monitoring data continues to increase, machine learning approaches have become increasingly attractive for supporting automated environmental monitoring, early warning systems, and data-driven decision-making. Their ability to process large datasets and identify hidden patterns has made ML a valuable tool for improving the accuracy and responsiveness of urban air quality assessment, particularly in metropolitan areas characterised by rapidly changing pollution dynamics. Among the repertoire of supervised learning algorithms, Support Vector Machine (SVM) has 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). The algorithm constructs an optimal decision boundary by maximizing the margin between different classes, thereby improving generalization performance on previously unseen data. Through the use of kernel functions, particularly the Radial

Basis Function (RBF) kernel, SVM can effectively capture complex non-linear relationships among multiple environmental variables without requiring explicit feature transformation. These characteristics make SVM well suited for air quality classification, where pollutant concentrations exhibit intricate interactions influenced by emission sources, meteorological conditions, and temporal variability. Furthermore, SVM is relatively robust to high-dimensional feature spaces and has been shown to achieve competitive predictive performance while maintaining computational efficiency, making it an attractive approach for developing reliable and scalable air quality monitoring systems (Zhu, Wang, Yang, Zhang, Zhang, Ren, Wu, & Ye, 2022; Adawiyah, & Iskandar, 2022).

Despite a growing body of literature applying ML to air quality prediction globally, several research gaps persist in the Indonesian context. First, most existing domestic studies employ single algorithms without rigorous hyperparameter optimisation, limiting the reliability of reported accuracy metrics (Saputra, & Firmansyah, 2025; Rybarczyk, & Zalakeviciute, 2018). 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 through the development and evaluation of an optimised SVM classification model for ISPU prediction in DKI Jakarta, grounded in a comprehensive, reproducible methodology. Specifically, this research: (1) constructs a preprocessing pipeline integrating missing value handling, label encoding, Min-Max normalisation, and stratified train-test splitting; (2) optimises SVM kernel parameters (C and γ) using GridSearchCV with StratifiedKFold 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.

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 recognised weakness in the regional ML-for-environment literature and provides a validated methodological template for researchers and practitioners across Indonesian cities. The remainder of the paper is organised 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.

2. Literature Review and Hypothesis/es Development

2.1 Air Pollution and Air Quality Index

Air pollution is one of the most significant environmental challenges affecting human health and sustainable urban development. It results from the emission of harmful substances originating from transportation, industrial activities, energy production, biomass burning, and natural processes (Karimi, Arif, & Noori, 2025). These pollutants are generally classified into primary pollutants, which are emitted directly into the atmosphere, and secondary pollutants, which are formed through chemical reactions involving primary emissions and atmospheric constituents (Yasin, Yasin, Iguala, Gelete, Tulu, & Kebede, 2025). In Indonesia, ambient air quality is evaluated using the Air Pollution Standard Index (ISPU), which is conceptually equivalent to the Air Quality Index (AQI) adopted in many countries. The ISPU integrates concentrations of major pollutants, including PM2.5, PM10, CO, SO_2, NO_2, and O_3, into a standardized index that represents potential health risks (Ramadhan, Hermawan, & Hudjimartsu, 2025; Pratama, Gudiato, Maulana, & Prayitno, 2025). Higher index

values indicate poorer air quality and greater health concerns, particularly for vulnerable populations such as children, older adults, and individuals with respiratory or cardiovascular diseases. By transforming complex environmental measurements into easily interpretable categories, the AQI provides an effective basis for public health communication, environmental policy implementation, and operational decision-making in air quality management. Consequently, accurate AQI classification has become an essential component of intelligent environmental monitoring systems (Talreja, Hegde, Kumar, & Chavali, 2025; McCarron, Semple, Braban, Swanson, Gillespie, & Price, 2023).

The increasing availability of environmental monitoring networks has enabled continuous collection of air quality data at high temporal resolution, creating opportunities for advanced data-driven analysis (Mahajan, Chung, Martinez, Olaya, Helbing, & Chen, 2022). However, the dynamic nature of atmospheric processes, influenced by meteorological conditions, emission sources, and seasonal variability, makes air quality assessment a complex task. Rather than relying solely on pollutant concentration thresholds, modern air quality management increasingly emphasizes predictive analytics to anticipate changes in pollution levels before they reach hazardous conditions. Consequently, AQI classification has evolved beyond a descriptive indicator into an important decision-support tool for environmental authorities, facilitating early warning systems, pollution mitigation strategies, and evidence-based urban planning (Khan, Amin, Khan, Jabeen, & Chaudhry, 2024); .

2.2 Urban Air Pollution and the AQI Framework

Urban air pollution arises from the interaction of primary emissions—directly released from combustion sources—and secondary pollutants formed through atmospheric chemical reactions (Zhang, Bocquet, Mallet, Seigneur, & Baklanov, 2012). 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 carrying distinct health implications. PM2.5 and PM10 are particularly critical given their deep respiratory penetration and associations with cardiovascular and pulmonary mortality (World, 2018). AQI frameworks such as ISPU aggregate multi-pollutant data into a single communicable index, enabling public health advisories calibrated to pollution severity (Kementerian, 2020).

AQI classification also plays a role in supporting evidence-based environmental management by providing a standardized representation of ambient air quality conditions. Accurate classification will help government agencies issue timely health warnings, implement pollution mitigation strategies, and ensure the effectiveness of environmental policies (Mahajan, Chung, Martinez, Olaya, Helbing, & Chen, 2022). Jakarta, a densely populated metropolitan area, experiences significant fluctuations in pollutant concentrations due to traffic emissions, industrial activity, and meteorological conditions. Reliable AQI prediction is crucial for reducing public exposure to hazardous air pollution. Therefore, applying machine learning techniques to classify AQI categories offers a practical approach to improving environmental monitoring systems by identifying complex and non-linear relationships among various pollutant variables and facilitating real-time automated decision-making support (Essamlali, Nhaila, & El, 2024).

Recent advances in environmental sensing technologies and open-data initiatives have significantly increased the availability of high-resolution air quality observations, creating new opportunities for data-driven environmental analysis. However, the large volume, high dimensionality, and dynamic nature of air quality data present considerable challenges for conventional analytical methods, particularly in accurately distinguishing between multiple AQI categories under rapidly changing urban conditions (Al-Thani & Isaifan, 2024). Machine learning-based classification methods provide an effective alternative by automatically extracting complex patterns from historical pollutant measurements without relying on explicit atmospheric assumptions. As a result, these approaches can improve the accuracy, consistency, and efficiency of AQI classification while supporting the

development of intelligent air quality monitoring systems capable of delivering timely information for public health protection and sustainable urban environmental management (Wan, Shackley, Doherty, Shi, Zhang, & Golding, 2020).

2.3 Machine Learning for Air Quality Classification

The application of ML to air quality prediction has evolved substantially since early comparative studies pitting regression models against neural networks for PM10 forecasting (Chaloulakou, Grivas, & Spyrellis, 2003). Subsequent advances have explored ensemble methods, deep learning architectures, and hybrid approaches integrating atmospheric physics with data-driven components (Zhu, Wang, Yang, Zhang, Zhang, Ren, Wu, & Ye, 2022). A comprehensive review by Zhang et al. (2012) documented the rapid evolution of real-time air quality forecasting frameworks, emphasising the role of data quality, feature selection, and model validation protocols in determining predictive reliability.

Within the classification paradigm, SVM has received sustained attention for air quality applications. Its theoretical foundation in statistical learning theory—particularly the structural risk minimisation principle—confers robustness advantages over empirical risk minimisation algorithms, especially under limited training data conditions (Hastie, Tibshirani, & Friedman, 2009). The kernel trick, which implicitly maps inputs into high-dimensional feature spaces, enables SVM to construct non-linear decision boundaries without explicit feature engineering—a critical capability given the complex multivariate interactions among atmospheric pollutants (Kumar, Gulia, Harrison, & Khare, 2017); (Jayadi, Handhayani, & Lauro, 2023); (Zou, Chen, Zhai, Fang, & Zheng, 2017).

2.4 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). Relative 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; Wen, Liu, Yao, Peng, Li, Hu, & Chi, 2019).

In the Indonesian context, (Yu, Li, Xie, & Wang, 2024) conducted a direct comparison of SVM and Naïve Bayes classifiers for air quality categorisation, demonstrating SVM superiority with an accuracy of 89.3% versus 76.4% for Naïve Bayes. (Saputra & Firmansyah, 2025) applied SVM alongside Decision Tree classifiers for urban air quality prediction, reporting that SVM performance was highly sensitive to kernel selection and regularisation parameter C, underscoring the importance of systematic hyperparameter optimisation. (Susatyono & Fitrianto, 2021) demonstrated that ML approaches, including SVM, outperformed statistical time-series models for categorical AQI prediction. (Fazle, Prodhan, & Islam, 2023) applied SVM for air quality data classification, emphasising the importance of multi-metric evaluation beyond accuracy alone.

2.5 Hyperparameter Optimization

The predictive performance of Support Vector Machine (SVM) models is highly dependent on the selection of appropriate hyperparameters. Unlike model parameters learned automatically during training, hyperparameters are specified before model fitting and directly influence the complexity and generalization capability of the classifier. For SVM with a Radial Basis Function (RBF) kernel, the penalty parameter (C) controls the trade-off between maximizing the decision margin and minimizing classification errors, whereas the gamma (γ) parameter determines the influence of individual training samples on the decision boundary. Improper hyperparameter selection may result in either underfitting or overfitting, leading to reduced predictive accuracy (Pannakkong, Thiwa-Anont, Singthong, Parthanadee, & Buddhakulsomsiri, 2022). Therefore, systematic optimization techniques such as GridSearchCV are widely employed to evaluate multiple hyperparameter combinations using

cross-validation. When combined with Stratified K-Fold cross-validation, GridSearchCV preserves class distributions across validation folds while providing a robust estimate of model performance. This optimization process improves model stability, enhances generalization ability, and facilitates the identification of the most suitable hyperparameter configuration for multi-class air quality classification (Hassan, Majeed, & Ahmad, 2025).

Recent advances in machine learning have demonstrated that hyperparameter optimization not only improves predictive accuracy but also enhances model robustness and reproducibility. An optimized hyperparameter configuration enables the classifier to construct decision boundaries that effectively capture complex patterns while minimizing the risk of poor generalization to unseen data (Hassan et al., 2025). In multi-class environmental classification problems, systematic optimization is particularly important because different pollutant categories often exhibit overlapping characteristics and unequal class distributions. Therefore, integrating hyperparameter tuning into the model development process has become a standard practice for producing reliable and unbiased classification models suitable for real-world environmental monitoring applications (Açıkkar & Altunkol, 2023).

2.6 Data Preprocessing for Air Quality Classification

Data preprocessing is a fundamental stage in machine learning because the quality of input data strongly influences model performance. Air quality datasets frequently contain missing values, duplicate observations, inconsistent categorical labels, and features measured on different numerical scales. These issues can reduce classification accuracy if not properly addressed before model training (Miah & Islam, 2024). Typical preprocessing procedures include missing value imputation to maintain dataset completeness, duplicate removal to eliminate redundant information, and label encoding to convert categorical target variables into numerical representations suitable for machine learning algorithms (Gressent, Malherbe, Colette, Rollin, & Scimia, 2020). Feature scaling is particularly important for Support Vector Machine because the algorithm relies on distance-based optimization. Min-Max normalization is commonly applied to transform predictor variables into a uniform range between 0 and 1 while preserving their relative distributions. This scaling prevents variables with larger numerical ranges from dominating the optimization process and contributes to faster convergence and improved classification performance. Comprehensive preprocessing therefore provides a consistent and reliable input representation that supports robust and accurate air quality classification (Maharana, Mondal, & Nemade, 2022).

Beyond improving data quality, preprocessing contributes to the computational efficiency and stability of machine learning algorithms. Standardized and well-prepared datasets reduce numerical instability during model optimization and enable the classifier to learn meaningful relationships among predictor variables (Hayward, Martin, Ferracci, Kazemimanesh, & Kumar, 2024). In air quality applications, preprocessing also helps mitigate the effects of measurement inconsistencies arising from sensor calibration, monitoring frequency, and environmental variability. As a result, a comprehensive preprocessing pipeline not only increases classification accuracy but also improves the reliability, scalability, and reproducibility of machine learning models developed for automated air quality assessment and real-time environmental decision support (Miah & Islam, 2024).

2.7 Identified Research Gaps

A synthesis of the literature reveals three principal gaps motivating 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—a critical omission given the class-imbalanced nature of AQI datasets (Ketu & Mishra, 2021). Second, preprocessing pipelines are frequently underspecified in published studies, preventing reproducibility assessment and methodological comparison across studies. Third, the specific context of Jakarta ISPU dataset from the Satu Data Jakarta portal remains underexplored with rigorous ML methodology, despite its public

availability and policy relevance. The present study addresses each of these gaps directly (Teguh, Oktaviyani, & Mempun, 2018; Putri, Handayani, & Rose, 2020).

2.8 Theoretical Framework

The theoretical framework of this research integrates two complementary perspectives. From the computational intelligence perspective, SVM is grounded in Vapnik-Chervonenkis (VC) theory, which formalises the relationship between model complexity, training error, and generalisation performance (Hastie, Tibshirani, & Friedman, 2009). The bias-variance trade-off operationalised through regularisation parameter C and kernel bandwidth γ is central to this framework. From the environmental informatics perspective, the study conceptualises 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 (Zhang, Bocquet, Mallet, Seigneur, & Baklanov, 2012).

2.9 Performance Evaluation

Performance evaluation is a critical stage in machine learning to assess the effectiveness and reliability of a classification model. It provides quantitative measures of how accurately the model predicts target classes and identifies potential weaknesses in classification performance. In multi-class classification problems, relying solely on overall accuracy may produce misleading conclusions, particularly when the dataset contains imbalanced class distributions. Therefore, multiple evaluation metrics are commonly employed to provide a more comprehensive assessment of model performance (Zhu, Gutierrez, Zhu, Ju, & Sarna, 2021). These metrics include accuracy, precision, recall, and F1-score, each offering different perspectives on the classifier's predictive capability. Accuracy measures the proportion of correctly classified instances, while precision evaluates the proportion of positive predictions that are correct. Recall reflects the model's ability to identify all instances belonging to a specific class, and the F1-score combines precision and recall into a single metric that balances both measures. Together, these indicators provide a robust evaluation framework for multi-class classification tasks (Maleki, Ovens, Gupta, Reinhold, Spatz, & Forghani, 2022).

In addition to numerical performance metrics, the confusion matrix is widely used to visualize classification outcomes by comparing predicted labels with actual class labels. This matrix provides detailed information regarding correctly classified observations and misclassification patterns for each category, enabling researchers to identify classes that are difficult to distinguish (Maleki et al., 2022). For imbalanced datasets, macro-averaged precision, recall, and F1-score are often preferred because they assign equal importance to every class regardless of sample size, thereby providing a more objective assessment of overall classification performance. A comprehensive evaluation using both statistical metrics and confusion matrix analysis ensures that the developed model is not only highly accurate but also reliable, robust, and capable of generalizing to unseen air quality data (Janiesch, Zschech, & Heinrich, 2021).

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 optimisation → 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 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, O_3) alongside an ISPU categorical label assigned according to KLHK's (2020) classification thresholds. All data processing and modelling was conducted on the Google Colaboratory cloud platform using Python 3 with the Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn libraries.

3.3 Data Preprocessing

The preprocessing pipeline comprised four sequential steps. First, missing value handling involved 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 any repeated observations arising from data export artefacts. Third, label encoding converted the categorical ISPU target variable into integer representations: Good = 0, Moderate = 2, Unhealthy = 3, Very Unhealthy = 4, Hazardous = 5. Fourth, Min-Max normalisation was applied to all six continuous predictor features to rescale values to the [0, 1] interval using the formula:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

Formula (1) shows the Min-Max normalization formula used in the preprocessing stage. It rescales each predictor variable to the range of 0–1 while maintaining the relative relationships among observations.

This normalisation step is essential to prevent features with larger absolute ranges (e.g., CO measured in µg/m³) from disproportionately influencing the SVM kernel distance calculations. 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 splits.

3.4 Model Development: Support Vector Machine

The SVM classifier was implemented using Scikit-learn SVC class. 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 (Villegas-Camacho, Francisco-Valencia, Alejo-Eleuterio, Granda-Gutiérrez, Martínez-Gallegos, & Villanueva-Vásquez, 2025; Tufail, Riggs, Tariq, & Sarwat, 2023). The multi-class extension employed the 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.

3.5 Hyperparameter Optimisation

Systematic hyperparameter optimisation was conducted using Scikit-learn GridSearchCV procedure. The parameter search space included regularisation 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 proportional class representation within each fold. The combination yielding the highest mean cross-validation accuracy was selected as the final model configuration. The random state was fixed at 42 to ensure reproducibility of the train-test split and fold assignments.

3.6 Model Evaluation

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 (Sensitivity): 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-average and Weighted-average Metrics: Providing unweighted and class-frequency-weighted aggregations, respectively.
6. Confusion Matrix: Visualising 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: macro-average treats all classes equally, while weighted-average accounts for class frequency, potentially masking poor minority class performance (Hastie, Tibshirani, & Friedman, 2009).

4. Results and Discussions

4.1 Dataset Characteristics and Class Distribution

The 1,825-record ISPU dataset exhibits a notably imbalanced class distribution. The Moderate category accounts for the substantial majority of observations (approximately 74.3% of the total dataset), reflecting Jakarta chronic moderate pollution baseline. The Good category constitutes approximately 12.8% of records, while the Unhealthy category accounts 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 accuracy exceeding 74% without any genuine predictive capability.

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

4.2 Hyperparameter Optimisation Results

Table 1 summarises the cross-validation performance across 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 shows GridSearchCV identified the optimal hyperparameter combination as C = 10 and γ = 0.1 for the RBF kernel, yielding a mean cross-validation accuracy of 95.3% (SD = 0.8%) across the five training folds. The finding is consistent with the general guidance that moderate regularisation (C = 10) balanced with an intermediate kernel bandwidth (γ = 0.1) prevents both underfitting and

overfitting in multi-class SVM applications with normalised features (Hastie, Tibshirani, & Friedman, 2009).

4.3 Classification Performance on the Test Set

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

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

Based on Table 2, the Moderate category achieved the highest performance (F1 = 98%), reflective of its dominant representation in the training set. The Unhealthy category also demonstrated robust performance (F1 = 93%), indicating that the model captured the characteristic pollutant concentration patterns differentiating unhealthy from moderate conditions. The Good category showed slightly lower recall (88%) relative to precision (84%), suggesting 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 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 tasks.

4.4 Comparative Analysis with Prior Studies

The 96.1% overall accuracy achieved in this study compares favourably with prior Indonesian SVM-based air quality classification studies. Xu, Lu, Cetiner, and Taciroglu (2021) reported 89.3% accuracy on a smaller dataset without systematic hyperparameter optimisation, while Pholsook, Ramjan, and Wipulanusat (2025) achieved 91.4% with SVM on a different regional dataset. The performance improvement in the present study is attributable to: (1) the application of GridSearchCV for systematic hyperparameter optimisation; (2) the use of stratified cross-validation ensuring representative fold composition; and (3) the comprehensive preprocessing pipeline minimising noise-induced model interference. These findings reinforce the conclusion of Garcia-Atutxa, Villanueva-Flores, Barrio, Sanchez-Villamil, Martínez-Más, and Bueno-Crespo (2025) that methodological rigour is a primary determinant of ML model performance, beyond algorithm selection alone.

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 milliseconds for batch prediction with Scikit-learn SVC implementation—is compatible with real-time data processing pipelines fed by IoT sensor networks (Schizas, Karras, Karras, & Sioutas, 2022; Badawy, Ramadan, & Hefny, 2025). 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 triggering public health communications when model predictions indicate imminent category escalation. However, operational deployment must be conditioned on resolution of the class imbalance limitation described above, particularly given the life-safety implications of failing to detect Very Unhealthy or Hazardous conditions.

5. Conclusions

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 normalisation, stratified train-test splitting, RBF kernel SVM training, and GridSearchCV hyperparameter optimisation, 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 structural class imbalance in the training data. The discrepancy between macro-average (F1 = 86%) and weighted-average (F1 = 96%) metrics highlights the hazard of relying solely on overall accuracy for imbalanced multi-class classification evaluation.

The methodological contribution of this research lies in its reproducible, end-to-end pipeline—from raw governmental open data through multi-metric model evaluation—applied 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 archipelago.

Acknowledgement

The authors gratefully acknowledge the Environmental Agency (*Dinas Lingkungan Hidup*) of DKI Jakarta Province for making the ISPU monitoring data publicly accessible through the Satu Data Jakarta portal. The authors also extend appreciation to the Institut Informatika dan Bisnis Darmajaya for institutional support throughout this research. This study received no external financial funding.

Author Contributions

HTUM conceptualized the study, collected and curated the data, developed the data preprocessing pipeline, trained and optimized the machine learning model, analyzed the results, prepared the original draft, and revised the manuscript. CF supervised the research, contributed to the study conceptualization, reviewed the methodology, critically reviewed the results and discussion, and approved the final version of the manuscript.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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

Al-Thani, H. G., & Isaifan, R. J. (2024). Policies and regulations for sustainable clean air: an overview. *Sustainable Strategies for Air Pollution Mitigation: Development, Economics, and Technologies, 409-437*. https://doi.org/10.1007/698_2024_1093

Açıkkar, M., & Altunkol, Y. (2023). A novel hybrid PSO-and GS-based hyperparameter optimization algorithm for support vector regression. *Neural Computing and Applications*, *35*(27), 19961-19977. https://doi.org/10.1007/s00521-023-08805-5

Badawy, M., Ramadan, N., & Hefny, H. A. (2025). Towards Intelligent Healthcare: A Big Data and Machine Learning Approach for Real-Time Predictive Analytics. *Biomedical Materials & Devices, 1-23*. https://doi.org/10.1007/s44174-025-00568-y

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

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16, 321–357*. https://doi.org/10.1613/jair.953

Du, P., Bai, X., Tan, K., Xue, Z., Samat, A., Xia, J., ... & Liu, W. (2020). Advances of four machine learning methods for spatial data handling: A review. *Journal of Geovisualization and Spatial Analysis*, *4*(1), 13. https://doi.org/10.1007/s41651-020-00048-5

Essamlali, I., Nhaila, H., & El Khaili, M. (2024). Supervised machine learning approaches for predicting key pollutants and for the sustainable enhancement of urban air quality: A systematic review. *Sustainability*, *16*(3), 976. https://doi.org/10.3390/su16030976

Fazle, A. B., Prodhan, R. K., & Islam, M. M. (2023). AI-Powered Predictive Failure Analysis In Pressure Vessels Using Real-Time Sensor Fusion: Enhancing Industrial Safety And Infrastructure Reliability. *American Journal of Scholarly Research and Innovation*, *2*(02), 102-134. https://doi.org/10.63125/wk278c34

Garcia-Atutxa, I., Villanueva-Flores, F., Barrio, E. D., Sanchez-Villamil, J. I., Martínez-Más, J., & Bueno-Crespo, A. (2025). Artificial intelligence for ovarian cancer diagnosis via ultrasound: a systematic review and quantitative assessment of model performance. *Frontiers in Artificial Intelligence, 8, 1649746*. https://doi.org/10.3389/frai.2025.1649746

Gressent, A., Malherbe, L., Colette, A., Rollin, H., & Scimia, R. (2020). Data fusion for air quality mapping using low-cost sensor observations: Feasibility and added-value. *Environment international, 143, 105965*. https://doi.org/10.1016/j.envint.2020.105965

Hassan, T., Majeed, M., & Ahmad, M. (2025). Optimizing SVM Performance through Combinatorial Hyperparameter Tuning and Model Selection. *International Journal Bioautomation*, *29*(2), 117.. https://doi.org/10.7546/ijba.2025.29.2.000981

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

Hayward, I., Martin, N. A., Ferracci, V., Kazemimanesh, M., & Kumar, P. (2024). Low-cost air quality sensors: biases, corrections and challenges in their comparability. *Atmosphere*, *15*(12), 1523. https://doi.org/10.3390/atmos15121523

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

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic markets*, *31*(3), 685-695. https://doi.org/10.1007/s12525-021-00475-2

Jayadi, B. V., Handhayani, T., & Lauro, M. D. (2023). Perbandingan KNN dan SVM untuk klasifikasi kualitas udara di Jakarta. *Jurnal Ilmu Komputer dan Sistem Informasi*, *11*(2). https://doi.org/10.24912/jiksi.v11i2.26006

Karimi, M. S., Arif, S., & Noori, B. (2025). Greenhouse gases and their role in air pollution and global warming. *International Journal of Biological, Physical and Chemical Studies*, *7*(1), 06-13. https://doi.org/10.32996/ijbpcs.2025.7.1.2

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

Khan, M. I., Amin, A., Khan, M. T., Jabeen, H., & Chaudhry, S. R. (2024). Unveiling the hidden hazards of smog: health implications and antibiotic resistance in perspective. *Aerobiologia*, *40*(3), 353-372. https://doi.org/10.1007/s10453-024-09833-x

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

Mahajan, S., Chung, M. K., Martinez, J., Olaya, Y., Helbing, D., & Chen, L. J. (2022). Translating citizen-generated air quality data into evidence for shaping policy. *Humanities and Social Sciences Communications*, *9*(1), 1-18. https://doi.org/10.1057/s41599-022-01135-2

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1), 91-99. https://doi.org/10.1016/j.gltp.2022.04.020

Maleki, F., Ovens, K., Gupta, R., Reinhold, C., Spatz, A., & Forghani, R. (2022). Generalizability of machine learning models: quantitative evaluation of three methodological pitfalls. *Radiology: Artificial Intelligence*, *5*(1). https://doi.org/10.1148/ryai.220028

McCarron, A., Semple, S., Braban, C. F., Swanson, V., Gillespie, C., & Price, H. D. (2023). Public engagement with air quality data: using health behaviour change theory to support exposure-minimising behaviours. *Journal of Exposure Science & Environmental Epidemiology*, *33*(3), 321-331. https://doi.org/10.1038/s41370-022-00449-2

Miah, M. R., & Islam, M. M. (2024). Data Preprocessing and Feature Engineering Strategies for Large-Scale Predictive Modeling Applications. *Review of Applied Science and Technology*, *3*(01), 263-302. https://doi.org/10.63125/tqqqed47

Pannakkong, W., Thiwa-Anont, K., Singthong, K., Parthanadee, P., & Buddhakulsomsiri, J. (2022). Hyperparameter tuning of machine learning algorithms using response surface methodology: a case study of ANN, SVM, and DBN. *Mathematical problems in engineering*, *2022*(1), 8513719. https://doi.org/10.1155/2022/8513719

Pholsook, T., Ramjan, S., & Wipulanusat, W. (2025). Enhancing airport services: data-driven analysis of passenger satisfaction and service quality in Southeast Asia. *Engineering Management in Production and Services*, *17*(2), .. https://doi.org/10.2478/emj-2025-0011

Pratama, A., Gudiato, C., Maulana, A., & Prayitno, W. (2025). SVM-Driven ISPU Prediction for Jakarta's Sustainable Future. *SISFOTENIKA*, *15*(1), 66-76. https://doi.org/10.30700/sisfotenika.v15i1.532

Putri, M. N. P., Handayani, A. S., & Rose, M. M. (2020). Sistem monitoring kualitas udara dengan platform web. *Explore IT: Jurnal Keilmuan dan Aplikasi Teknik Informatika*, *12*(2), 54-61. https://doi.org/10.35891/explorit.v12i2.2262

Ramadhan, A., Hermawan, E., & Hudjimartsu, S. A. (2025). Spatial Distribution Patterns of Urban Growth, Dasymetric Mapping, and Air Pollution Standard Index (ISPU) in Bogor Regency (2018-2023). *Journal of Applied Informatics and Computing*, *9*(4), 1477-1489. https://doi.org/10.30871/jaic.v9i4.9862

Rybarczyk, Y., & Zalakeviciute, R. (2018). Machine learning approaches for outdoor air quality modelling: A systematic review. *Applied Sciences*, *8*(12), 2570. https://doi.org/10.3390/app8122570

Saputra, A. D. M., & Firmansyah, H. (2025). Implementasi Algoritma Decision Tree untuk Memprediksi Kualitas Udara dan Polusi dengan RapidMiner. *sudo Jurnal Teknik Informatika*, *4*(3), 212-219. https://doi.org/10.56211/sudo.v4i3.1145

Schizas, N., Karras, A., Karras, C., & Sioutas, S. (2022). TinyML for ultra-low power AI and large scale IoT deployments: A systematic review. *Future Internet*, *14*(12), 363. https://doi.org/10.3390/fi14120363

Susatyono, J. D., & Fitrianto, Y. (2021). Sistem monitoring kualitas udara dan otomatisasi pemberian pakan ayam berbasis IoT. *Krea-TIF: Jurnal Teknik Informatika*, *9*(2), 1-10. https://doi.org/10.32832/kreatif.v9i2.5650

Talreja, N., Hegde, C., Kumar, E. M., & Chavali, M. (2025). Emerging environmental contaminants: sources, consequences and future challenges. *Green technologies for industrial contaminants, 119-149*. https://doi.org/10.1002/9781394159390.ch5

Teguh, R., Oktaviyani, E. D., & Mempun, K. A. (2018). Rancang bangun desain internet of things untuk pemantauan kualitas udara pada studi kasus polusi Udara. *Jurnal Teknologi Informasi: Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika*, *12*(2), 47-58. https://doi.org/10.47111/jti.v12i2.532

Tufail, S., Riggs, H., Tariq, M., & Sarwat, A. I. (2023). Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms. *Electronics*, *12*(8), 1789. https://doi.org/10.3390/electronics12081789

Villegas-Camacho, O., Francisco-Valencia, I., Alejo-Eleuterio, R., Granda-Gutiérrez, E. E., Martínez-Gallegos, S., & Villanueva-Vásquez, D. (2025). FTIR-based microplastic classification: A comprehensive study on normalization and ML techniques. *Recycling*, *10*(2), 46. https://doi.org/10.3390/recycling10020046

Wan, K., Shackley, S., Doherty, R. M., Shi, Z., Zhang, P., & Golding, N. (2020). Science-policy interplay on air pollution governance in China. *Environmental Science & Policy, 107, 150-157*. https://doi.org/10.1016/j.envsci.2020.03.003

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*.

Xu, Y., Lu, X., Cetiner, B., & Taciroglu, E. (2021). Real-time regional seismic damage assessment framework based on long short-term memory neural network. *Computer-Aided Civil and Infrastructure Engineering*, *36*(4), 504-521. https://doi.org/10.1111/mice.12628

Yasin, K. H., Yasin, M. I., Iguala, A. D., Gelete, T. B., Tulu, D., & Kebede, E. (2025). Predictive machine learning and geospatial modeling reveal PM10 hotspots and guide targeted air pollution interventions in Addis Ababa, Ethiopia. *Discover Applied Sciences*, *7*(4), 263. https://doi.org/10.1007/s42452-025-06723-w

Yu, K., Li, W., Xie, W., & Wang, L. (2024). A hybrid feature-selection method based on mRMR and binary differential evolution for gene selection. *Processes*, *12*(2), 313. https://doi.org/10.3390/pr12020313

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

Zhu, M., Gutierrez, M., Zhu, H., Ju, J. W., & Sarna, S. (2021). Performance Evaluation Indicator (PEI): A new paradigm to evaluate the competence of machine learning classifiers in predicting rockmass conditions. *Advanced Engineering Informatics, 47, 101232*. https://doi.org/10.1016/j.aei.2020.101232

Zou, B., Chen, J., Zhai, L., Fang, X., & Zheng, Z. (2017). Satellite based mapping of ground PM2.5 concentration using generalized additive modeling. *Remote Sensing*, *9*(1), 1. https://doi.org/10.3390/rs9010001

 DOI:10.52909/jti.v6i1.324