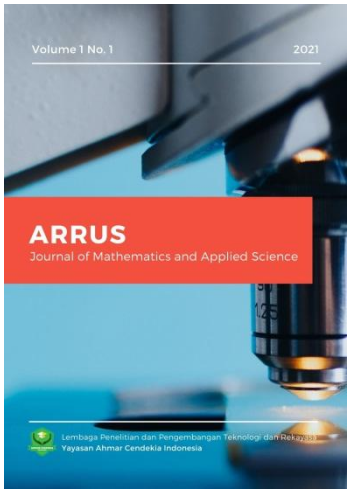
https://doi.org/10.35877/mathscience5581

OPEN ACCESS

ISSN : 2776-7922 (Print) / 2807-3037 (Online)



*Corresponding author: Zulkifli Rais, Department of Statistics, Universitas Negeri Makassar, Indonesia

E-mail: zulkifli.rais89@unm.ac.id

RESEARCH ARTICLE

Comparing SMOTE, Cost-Sensitive, and Fuzzy SVM for Day-Ahead Air Quality Category Forecasting in Jakarta

Zulkifli Rais*& Agung Tri Utomo

Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Makassar, Makassar, Indonesia.

Abstract: This study compares three strategies for handling class imbalance in multiclass support vector machines (SVM) for forecasting the next-day category of the Indonesian Air Pollution Standard Index (ISPU): SMOTE combined with SVM, a cost-sensitive SVM, and a fuzzy SVM for class imbalance learning (FSVM-CIL), with a standard SVM as the baseline. The data are 7,092 daily records from five air quality monitoring stations in Jakarta for 2021–2025, merged with NASA POWER meteorology. Predictors contain only lagged pollutant, meteorological, and calendar information, so no information from the target day enters the model. The Moderate category covers 76.2% of station-days, while the rarest category differs across stations. Models were tuned by time-series cross-validation, tested on 2024 and 2025, and compared with the Friedman test over ten station-year blocks. FSVM-CIL achieved the highest mean macro-F1 (0.521), balanced accuracy (0.601), and G-mean (0.502), against 0.445, 0.458, and 0.175 for the standard SVM, and raised the recall of Unhealthy days from 0.18 to 0.44. The differences were not statistically significant ($\chi^2 = 5.67$, $p = 0.129$), and the best model varied by station. Imbalance handling improves the detection of rare air quality categories, but the strategy should be chosen by local validation.

Keywords: Air Quality Index, Class Imbalance, Fuzzy Support Vector Machine, SMOTE, Time-Series Classification

1. Introduction

Air pollution in Jakarta is a major public health problem that imposes a substantial burden of disease and economic cost (Syuhada et al., 2023). Fine particulate matter ($PM_{2.5}$) concentrations in the city rise sharply during the dry season, particularly from May to August, when high temperatures, low humidity, weak winds, and scarce rainfall limit pollutant dispersion (Firdaus & Permadi, 2024). The interaction between emissions and meteorology causes daily air quality to fluctuate and makes it difficult to anticipate (Handhayani, 2023). Similar concerns arise in other Indonesian cities: air pollution in Makassar has been mapped with clustering methods (Annas et al., 2022) and linked to green space and CO_2 reduction (Ruliana et al., 2025), and Indonesian provinces have been grouped by their environmental pollution cases (Muttaqim et al., 2023).

In Indonesia, air quality is communicated to the public through the Air Pollution Standard Index (*Indeks Standar Pencemar Udara*, ISPU), regulated by Minister of Environment and Forestry Regulation No. P.14/MENLHK/SETJEN/KUM.1/7/2020 (Ministry of

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5581

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Environment and Forestry, 2020). This regulation added $PM_{2.5}$ to the index and defines five categories: Good (0–50), Moderate (51–100), Unhealthy (101–200), Very Unhealthy (201–300), and Hazardous (>300). These categories are the basis of activity advice for the public and for sensitive groups. The most valuable information for an early warning system is therefore the ISPU category of the next day, not the category of the current day, which can already be computed directly from the monitors.

Several studies have applied machine learning, particularly SVM, to classify ISPU categories in Jakarta (Ridho & Mahalisa, 2023; Jayadi et al., 2024; Salsabila & Utami, 2025), and support vector regression has been used to forecast air quality index values in Makassar (Rahmat et al., 2023). The classification studies share two limitations. First, they use pollutant concentrations measured on the same day as the target category as predictors. Because the ISPU category is a deterministic function of the maximum pollutant sub-index, such models merely relearn the category breakpoints. This is reflected in near-perfect accuracy, for example 99% for a decision tree on Jakarta ISPU data for 2010–2021 (Jayadi et al., 2024). Second, class imbalance is usually handled only with SMOTE and assessed with overall accuracy, which is biased towards the majority class (He & Garcia, 2009).

Strategies for handling class imbalance in SVM fall into three groups. Data-level methods rebalance the class distribution before training, for example SMOTE, which generates synthetic minority samples (Chawla et al., 2002). Algorithm-level, or cost-sensitive, methods assign larger misclassification penalties to minority classes (Veropoulos et al., 1999). The fuzzy SVM for class imbalance learning (FSVM-CIL) assigns each sample a membership value based on its class and its distance to the class centre, which also reduces the influence of outliers and noise (Batuwita & Palade, 2010). This property is relevant because monitoring-station data are prone to sensor faults, maintenance gaps, and missing values. To the best of our knowledge, these three approaches have not been systematically compared for forecasting ISPU categories in Indonesia.

This study therefore compares a standard SVM, SMOTE+SVM, a cost-sensitive SVM, and FSVM-CIL for forecasting the ISPU category one day ahead (D+1) at five monitoring stations (*Stasiun Pemantau Kualitas Udara*, SPKU) in Jakarta. The contributions of this study are as follows:

(1). The ISPU classification task is reformulated as leakage-free D+1 forecasting, with lagged pollutant and meteorological values as predictors.
(2). Three imbalance-handling strategies for SVM are compared using metrics suited to imbalanced data: macro-F1, balanced accuracy, G-mean, and the Matthews correlation coefficient (MCC).
(3). The consistency of the model ranking is tested on five stations with different imbalance patterns using the Friedman test and the Nemenyi post-hoc test (Demšar, 2006).
(4). A reproducible analysis pipeline based on open data from the Jakarta open data portal (Satu Data Jakarta) is provided.

2. Literature Review

2.1. *Air Pollution Standard Index (ISPU)*

ISPU is a dimensionless number that describes ambient air quality at a given location. Under Regulation P.14/2020, the sub-index of each pollutant is computed by linear interpolation between two concentration breakpoints:

$$I = \frac{I_a - I_b}{X_a - X_b}(X_x - X_b) + I_b \qquad (1)$$

where I is the ISPU sub-index of the pollutant, I_a and I_b are the upper and lower ISPU limits, X_a and X_b are the upper and lower ambient concentration limits, and X_x is the measured concentration. The daily ISPU of a station is the highest sub-index, and the pollutant that produces it is called the critical parameter. Equation (1) shows that the ISPU category is a

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5581

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

deterministic function of same-day concentrations. Consequently, a model that uses same-day concentrations as predictors does not forecast; it reconstructs the categorisation rule.

2.2. *Multiclass Support Vector Machine*

SVM seeks the separating hyperplane with the maximum margin between two classes (Cortes & Vapnik, 1995). For data that cannot be separated perfectly, the soft-margin formulation is used:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad y_i(\mathbf{w}^\top \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \qquad (2)$$

The parameter C controls the trade-off between margin width and training error. Non-linear relationships are handled with a kernel, such as the radial basis function (RBF) kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$. The multiclass case is solved with the one-versus-one strategy, which builds $K(K-1)/2$ binary classifiers and has proven competitive with other strategies (Hsu & Lin, 2002). On imbalanced data, a standard SVM tends to skew the hyperplane towards the minority class, so minority-class recall is low even when overall accuracy appears high (He & Garcia, 2009; Batuwita & Palade, 2010). In Indonesia, SVM and its regression variant have been applied to health classification (Fahmuddin S et al., 2025), student graduation classification (Ruliana et al., 2024), and commodity price forecasting (Ahmar et al., 2025).

2.3. *SMOTE*

The Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic minority samples by interpolating between a minority sample and one of its k nearest neighbours (Chawla et al., 2002):

$$\mathbf{x}_{\text{new}} = \mathbf{x}_i + \lambda\,(\mathbf{x}_{zi} - \mathbf{x}_i), \qquad \lambda \sim U(0,1) \qquad (3)$$

SMOTE enlarges the decision region of the minority class, but it may generate samples in regions where classes overlap. Variants such as SMOTE-ENN remove ambiguous samples after over-sampling (Batista et al., 2004).

2.4. *Cost-Sensitive SVM*

A cost-sensitive SVM assigns a different penalty parameter to each class, so that errors on minority classes are more costly (Veropoulos et al., 1999):

$$\min \frac{1}{2} \| \mathbf{w} \|^2 + \sum_{k=1}^{K} C_k \sum_{i:\, y_i = k} \xi_i\,, \qquad C_k = C \cdot \frac{n}{K\, n_k} \qquad (4)$$

where n is the number of training samples and n_k is the number of samples in class k. These inverse-frequency weights serve as a starting point and can also be treated as tunable hyper-parameters.

2.5. *Fuzzy SVM and FSVM-CIL*

The fuzzy SVM (FSVM) assigns each sample a membership value $s_i \in (0, 1]$, so that less representative samples, such as outliers, contribute less to the objective function (Lin & Wang, 2002):

$$\min \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{n} s_i\, \xi_i \qquad (5)$$

FSVM-CIL extends FSVM to imbalanced data by defining the membership as the product of a class factor and a distance factor (Batuwita & Palade, 2010):

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5581

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

$$s_i = r_{y_i} \cdot f(\mathbf{x}_i), \qquad f_{\text{lin}}(\mathbf{x}_i) = 1 - \frac{d_i^{\text{cen}}}{\max_j d_j^{\text{cen}} + \Delta}, \qquad f_{\exp}(\mathbf{x}_i) = \frac{2}{1 + \exp\left(\beta\, d_i^{\text{cen}}\right)} \qquad (6)$$

where d_i^{cen} is the distance of a sample to its class centre, Δ is a small constant, and $\beta \in [0, 1]$ is a decay parameter. In the original binary formulation, the minority class receives $r = 1$ and the majority class receives $r = n_{\text{min}}/n_{\text{maj}}$. This study extends it to the multiclass case with $r_k = n_{\text{min}}/n_k$, so the smallest class has weight 1 and other classes are weighted in proportion to their scarcity. Because the FSVM objective is identical to a sample-weighted SVM, FSVM-CIL can be implemented by passing s_i as sample weights to a standard SVM library.

2.6. *Related Studies*

Table 1 summarises the air quality classification studies most closely related to this work. Jayadi et al. (2024) reported 1,155 Good, 3,087 Moderate, and 141 Unhealthy days in the 2010–2021 data. This distribution differs from the data after $PM_{2.5}$ entered the ISPU calculation: the Jakarta Environmental Agency reported that Unhealthy days reached 43% in January–October 2021, whereas Good days accounted for only 1% (Dinas Lingkungan Hidup Provinsi DKI Jakarta, 2021). The minority class is therefore not fixed, and imbalance-handling strategies need to be tested under several distribution patterns.

Table 1. Related studies on air quality category classification

Study	Data	Predictors	Methods	Imbalance handling	Remarks
Salsabila & Utami (2025)	Jakarta ISPU and satellite imagery	Multi-source	SVM	SMOTE-Tomek Link	Classification and spatial mapping
Jayadi et al. (2024)	Jakarta ISPU 2010–2021 (4,383 samples)	Same-day PM_{10}, CO, SO_2, O_3, NO_2	SVM, NB, LR, DT, AdaBoost	Not specific	DT accuracy 99%
Saminathan & Malathy (2023)	12 Beijing stations, 2013–2017	Pollutants and meteorology	LR, SVM, RF, XGBoost, MLP	SMOTE	RF-SMOTE best
Rahmat et al. (2023)	Makassar air quality index	Past pollutant series	SVR (regression)	Not applicable	Forecasts index values, not categories
Ridho & Mahalisa (2023)	Jakarta ISPU, pandemic period	Same-day pollutants	SVM	None	—
This study	5 Jakarta SPKU, 2021–2025, and meteorology	Lagged pollutants and meteorology (D+1)	SVM, SMOTE+SVM, CS-SVM, FSVM-CIL	Three strategies compared	Temporal validation, Friedman test

Note: SVR = support vector regression; NB = naive Bayes; LR = logistic regression; DT = decision tree; RF = random forest; MLP = multilayer perceptron; CS-SVM = cost-sensitive SVM.

3. Research Method and Materials

3.1. *Data*

Air quality data are daily ISPU values from the five SPKU operated by the Jakarta Environmental Agency and published on Satu Data Jakarta (Dinas Lingkungan Hidup Provinsi DKI Jakarta, 2023a, 2023b, 2026): DKI1 Bundaran HI, DKI2 Kelapa Gading, DKI3 Jagakarsa, DKI4 Lubang Buaya, and DKI5 Kebon Jeruk. Each record contains the date, station, sub-indices of PM_{10}, $PM_{2.5}$, SO_2, CO, O_3, and NO_2, the maximum value, the critical parameter, and the category. The study period is January 2021 to November 2025, that is, after $PM_{2.5}$ entered the ISPU calculation. Station-level records for January–November 2022 are not available because the 2022 dataset reports only the station with the highest daily index (Dinas Lingkungan Hidup Provinsi DKI Jakarta, 2023c), and December 2023 and December

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5581

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

2025 had not been published at the time of retrieval (8 November 2026). The final dataset contains 7,150 station-days, of which 7,092 carry a valid category.

Daily meteorological data were obtained from the NASA POWER daily point API (NASA Langley Research Center, 2026) for central Jakarta (6.18°S, 106.83°E): mean air temperature at 2 m (°C), relative humidity at 2 m (%), corrected precipitation (mm/day), wind speed at 2 m (m/s), wind direction at 2 m, encoded as sine and cosine, and all-sky surface shortwave radiation (kWh/m²/day). The series covers December 2020 to December 2025 without gaps. Because the POWER grid (0.5° × 0.625°) contains all five stations, the same meteorological series is used for every station.

3.2. *Pre-processing*

Non-numeric pollutant entries (for example "---" or "N/A") were set as missing. Days with a missing target category were removed, and the target was never imputed. Missing predictor values were filled by linear interpolation for gaps of up to three consecutive days. Pollutant sub-indices that were still missing were replaced by the median of the training data of each fold, and a binary missing-value indicator was added for each affected variable. Station-years in which fewer than 75% of calendar days carry a valid category were excluded as target years, following the data-completeness threshold of Regulation P.14/2020; this removes December 2022, which covers only 31 days. Predictors were standardised with z-scores whose means and standard deviations were computed from the training data of each fold only.

3.3. *Target Variable*

The target is the ISPU category one day ahead at station s:

$$y_(s, t+1) \in \{\text{Good},\ \text{Moderate},\ \text{Unhealthy}^{+}\} \qquad (7)$$

The Unhealthy⁺ class merges Unhealthy, Very Unhealthy, and Hazardous, because the last two categories occurred on only four station-days in the study period. The merger also matches early-warning needs, since all three categories trigger control measures under Regulation P.14/2020.

3.4. *Feature Engineering and Leakage Prevention*

All predictors use only information available by the end of day t. The features are (1) the sub-indices of the six pollutants and the maximum ISPU at t, $t-1$, and $t-2$; (2) the mean and maximum ISPU over the last seven days; (3) the ISPU category at t in ordinal coding; (4) meteorological values at t and $t-1$ and the number of consecutive dry days up to t; and (5) calendar features for $t+1$ (day of week, month as sine and cosine, and a weekend indicator), which are known in advance. This gives 42 predictors. The maximum value, critical parameter, and category of day $t+1$ are never used as predictors. As an additional analysis, a perfect-prognosis scenario adds the meteorology of day $t+1$ to indicate the upper bound of performance if perfect weather forecasts were available.

3.5. *Models*

All four models use the RBF kernel and the one-versus-one strategy (Table 2). In M3, class centres and distances d_i^{cen} are computed in the standardised feature space from training data only, and s_i is passed as the sample weight of the SVM, which is mathematically equivalent to the FSVM objective.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5581

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Table 2: Models and hyper-parameter search space

Code	Model	Imbalance handling	Hyper-parameters searched
M0	Standard SVM	None	$C \in \{0.1, 1, 10, 100\}$; $\gamma \in \{0.001, 0.01, 0.1, \text{scale}\}$
M1	SMOTE+SVM	SMOTE on the training data of each fold until all classes are balanced	C, γ; k neighbours ∈ {3, 5, 7}
M2	Cost-sensitive SVM	Class weights $C_k = C \cdot n/(K \cdot n_k)$	C, γ
M3	FSVM-CIL	Membership $s_i = r_k \cdot f(x_i)$, $r_k = n_{min}/n_k$	C, γ; f ∈ {lin, exp}; β ∈ {0.1, 0.5, 1}

Note: "scale" sets $\gamma = 1/(p \cdot \text{Var}(X))$, where p is the number of features.

3.6. *Validation Scheme*

Time-series data must not be shuffled, because shuffling leaks future patterns into the training data (Bergmeir & Benítez, 2012). The outer evaluation uses an annual rolling origin. In the first fold, the models are trained on all data before 2024 (2021 and January–November 2023) and tested on 2024. In the second fold, they are trained on all data up to 2024 and tested on January–November 2025. Within each training set, hyper-parameters are selected by five-fold expanding-window time-series cross-validation that maximises macro-F1. Imputation, standardisation, SMOTE, and membership weights are recomputed in every fold to avoid information leakage. Five stations and two test years give N = 10 evaluation blocks. Figure 1 summarises the research workflow.

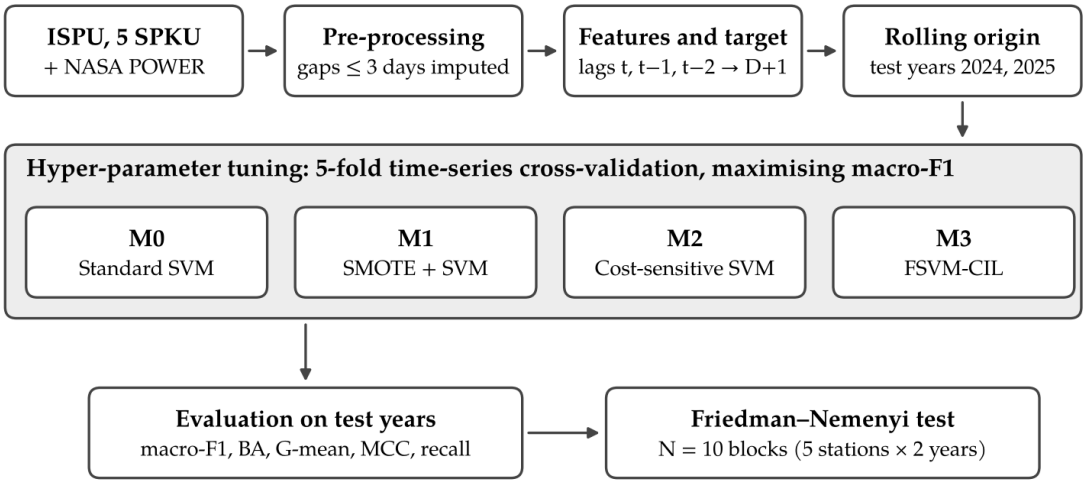


Figure 1: Research workflow

3.7. *Evaluation Metrics*

Accuracy is not used as the main metric because it is biased towards the majority class. With Rec_k the proportion of class-k samples predicted correctly, Prec_k the precision of class k, and $K = 3$, the metrics are:

$$\text{Macro-F1} = \frac{1}{K}\sum_{k=1}^{K}\frac{2\,\text{Prec}_k\,\text{Rec}_k}{\text{Prec}_k + \text{Rec}_k}, \quad \text{BA} = \frac{1}{K}\sum_{k=1}^{K}\text{Rec}_k, \quad \text{G-mean} = \left(\prod_{k=1}^{K}\text{Rec}_k\right)^{1/K} \qquad (8)$$

The multiclass MCC, which accounts for all cells of the confusion matrix, is also reported (Chicco & Jurman, 2020). The recall of the minority class at each station is reported separately because it matters most for early warning.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5581

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

3.8. *Statistical Test*

Differences among the four models were tested with the Friedman test over N = 10 blocks, using macro-F1 as the measure (Demšar, 2006). If the null hypothesis is rejected at $\alpha = 0.05$, the Nemenyi post-hoc test is applied with the critical difference:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}, \qquad k = 4,\ N = 10,\ q_{0.05} = 2.569 \ \Rightarrow\ CD \approx 1.48 \qquad (9)$$

Two models differ significantly if the difference in their average ranks exceeds CD. The analysis was carried out in Python 3 with pandas, scikit-learn (Pedregosa et al., 2011), and SciPy; SMOTE was implemented directly following the algorithm of Chawla et al. (2002).

4. Results and Discussion

4.1. *Data Description*

Table 3 shows that the Moderate category dominates every station, accounting for 76.2% of all station-days. The two remaining classes are almost equally rare overall (Good 12.2%, Unhealthy⁺ 11.6%), but the minority class differs between stations. At DKI1 Bundaran HI, Unhealthy⁺ is the rarest class (5.4%), whereas at DKI3 Jagakarsa the rarest class is Good (5.6%). DKI4 Lubang Buaya has the most frequent Unhealthy⁺ days (26.1%) and the lowest imbalance ratio together with DKI5 Kebon Jeruk. The imbalance ratio between the largest and smallest class ranges from 5.8 to 15.6, which provides five distinct imbalance scenarios for comparing the models. After lag construction and the completeness rule, 6,838 station-days remained for modelling, of which 3,436 form the two test years (1,793 in 2024 and 1,643 in 2025).

Table 3: Distribution of daily ISPU categories by station, January 2021–November 2025

Station	n (days)	Good (%)	Moderate (%)	Unhealthy⁺ (%)	Imbalance ratio
DKI1 Bundaran HI	1,424	15.0	79.6	5.4	14.7
DKI2 Kelapa Gading	1,422	10.1	81.3	8.6	9.4
DKI3 Jagakarsa	1,420	5.6	88.1	6.3	15.6
DKI4 Lubang Buaya	1,406	10.5	63.4	26.1	6.0
DKI5 Kebon Jeruk	1,420	19.9	68.2	11.8	5.8
All stations	7,092	12.2	76.2	11.6	6.6

Note: Imbalance ratio = size of the largest class / size of the smallest class. Station-level data for January–November 2022, December 2023, and December 2025 are not available. Source: computed from Satu Data Jakarta.

4.2. *Model Performance*

Table 4 shows that all three imbalance-handling strategies outperformed the standard SVM on every imbalance-sensitive metric. FSVM-CIL (M3) achieved the highest mean macro-F1 (0.521), balanced accuracy (0.601), G-mean (0.502), and MCC (0.300), followed by the cost-sensitive SVM (M2). Relative to M0, M3 increased macro-F1 by 0.075 and G-mean by 0.327. The standard SVM had the highest accuracy (0.763) but the lowest balanced accuracy (0.458) and a G-mean of only 0.175, because it predicted 86% of test days as Moderate. This pattern confirms that accuracy is misleading on imbalanced data.

M3 outperformed M0 in seven of the ten blocks and was the best model in five blocks. The Friedman test, however, did not reject equal performance ($\chi^2(3) = 5.67$, $p = 0.129$; Iman–Davenport $F(3, 27) = 2.10$, $p = 0.124$). M3 had the best average rank (1.90), followed by M2 (2.15), M0 (2.95), and M1 (3.00); the largest rank difference, between M1 and M3 (1.10), is below the Nemenyi critical difference of 1.48. With only ten blocks the test has limited power, so the differences should be read as consistent tendencies rather than confirmed effects.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5581

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Table 4: Mean (standard deviation) model performance over ten test blocks and Friedman average ranks

Model	Macro-F1	Balanced accuracy	G-mean	MCC	Accuracy	Mean rank
M0 Standard SVM	0.445 (0.092)	0.458 (0.086)	0.175 (0.240)	0.226 (0.133)	0.763 (0.098)	2.95
M1 SMOTE+SVM	0.478 (0.088)	0.561 (0.126)	0.451 (0.251)	0.263 (0.127)	0.658 (0.148)	3.00
M2 Cost-sensitive SVM	0.508 (0.066)	0.574 (0.108)	0.470 (0.216)	0.284 (0.119)	0.683 (0.123)	2.15
M3 FSVM-CIL	0.521 (0.072)	0.601 (0.096)	0.502 (0.226)	0.300 (0.109)	0.684 (0.114)	1.90

Note: Mean rank is the Friedman average rank on macro-F1 (1 = best). Friedman $\chi^2(3) = 5.67$, $p = 0.129$; Iman–Davenport $F(3, 27) = 2.10$, $p = 0.124$; Nemenyi CD ($\alpha = 0.05$, $k = 4$, $N = 10$) = 1.48. No pair of models differs significantly.

4.3. *Performance by Station and Minority Class*

Table 5 shows that the best model depends on the station. M3 had the highest macro-F1 at DKI2, DKI4, and DKI5, and M2 at DKI1, whereas at DKI3 Jagakarsa no strategy improved on the standard SVM (0.434) and SMOTE performed worst (0.359). DKI3 is the most imbalanced station (Table 3): each rare class has only 34 test days, and the training data contain few minority examples from which SMOTE can interpolate. The largest gain occurred at DKI5 in 2024, where macro-F1 rose from 0.263 (M0) to 0.582 (M3). Across stations, M3 raised the recall of Good days from 0.20–0.42 to 0.55–0.85 and the recall of Unhealthy$^+$ days from 0.00–0.49 to 0.25–0.61. At DKI1 Bundaran HI, the standard SVM did not detect a single Unhealthy$^+$ day.

Table 5: Macro-F1 by station and recall of the two rare classes for M0 and M3, mean of the two test years

Station	M0	M1	M2	M3	Recall Good (M0 → M3)	Recall Unhealthy$^+$ (M0 → M3)
DKI1 Bundaran HI	0.442	0.427	0.523	0.496	0.42 → 0.62	0.00 → 0.30
DKI2 Kelapa Gading	0.477	0.569	0.568	0.579	0.40 → 0.77	0.05 → 0.29
DKI3 Jagakarsa	0.434	0.359	0.423	0.422	0.36 → 0.55	0.04 → 0.25
DKI4 Lubang Buaya	0.461	0.545	0.549	0.565	0.28 → 0.85	0.49 → 0.61
DKI5 Kebon Jeruk	0.413	0.490	0.479	0.542	0.20 → 0.78	0.17 → 0.56

Note: Columns M0–M3 report mean macro-F1. Test-set counts (Good/Moderate/Unhealthy$^+$): DKI1 107/542/44; DKI2 80/541/68; DKI3 34/619/34; DKI4 110/490/76; DKI5 111/466/114.

Pooled over all blocks, M3 detected 149 of 336 Unhealthy$^+$ days (44%) and 324 of 442 Good days (73%), against 59 (18%) and 162 (37%) for M0. The gain came at the cost of more false alarms: M3 predicted 419 Moderate days as Unhealthy$^+$, so the precision of the Unhealthy$^+$ class remained low (0.26 for M3 and 0.29 for M0). Severe errors between Good and Unhealthy$^+$ were rare for all models (3–15 of 3,436 test days); most errors occurred between adjacent classes.

4.4. *Discussion*

Three mechanisms help interpret the results. First, SMOTE generates synthetic samples between neighbouring minority samples. In ISPU data, the boundary between Moderate and Unhealthy$^+$ is a sharp threshold at a sub-index of 100, so synthetic samples near the threshold may overlap with the Moderate class. This is consistent with M1 producing the most Moderate days predicted as Unhealthy$^+$ (511), the lowest Unhealthy$^+$ precision (0.25), and the weakest result at DKI3, where very few minority samples are available for interpolation.

Second, the cost-sensitive SVM shifts the decision boundary without changing the data. It adds no samples and gives all samples of a class the same weight, and it achieved the highest pooled Unhealthy$^+$ recall (0.53). Third, FSVM-CIL lowers the weight of samples far from their class centre. In SPKU data, such samples may come from sensor faults or unusual episodes. This down-weighting gave M3 the best balance between classes, with the highest G-mean and the highest recall of Good days, although its Unhealthy$^+$ recall (0.44) was below that of M1 and M2. Extreme pollution episodes are informative samples that FSVM-CIL also down-weights, which may explain this trade-off. M3 selected the largest penalty ($C = 100$) in

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

every block, because membership values below one reduce the effective penalty; a wider search range may improve it further.

The different imbalance patterns across stations have a practical implication. A single strategy is not optimal for every location: M3 was best at three stations, M2 at one, and the standard SVM at the most imbalanced station. An early warning system should therefore select the strategy for each station by local validation. The perfect-prognosis scenario, which adds next-day meteorology, raised the mean macro-F1 of M3 only from 0.521 to 0.538 and that of M0 from 0.445 to 0.466. At the spatial resolution used here, next-day weather information adds limited value beyond lagged pollutant and meteorological data.

This study has several limitations. Meteorological data come from a single reanalysis-based grid point, so spatial variation of weather among the SPKU is not captured. Station-level data for most of 2022 are unavailable, which shortens the training series, and ten evaluation blocks give the Friedman test limited power. The precision of the Unhealthy$^+$ class remains low, so false alarms are frequent. Merging Very Unhealthy and Hazardous with Unhealthy removes the ability to distinguish extreme episodes. The ordinal nature of ISPU categories is not yet exploited; ordinal classification or an asymmetric cost matrix is a natural extension. Only SVM variants were compared; gradient-boosting ensembles, which have performed well in multiclass health classification (Rais et al., 2025) and in rainfall classification in Makassar (Hafid et al., 2025; Rais et al., 2026), are natural benchmarks for future work. Finally, the low-cost sensors now widely deployed in Jakarta were not used because they require calibration against reference stations.

5. Conclusion

This study reformulated ISPU classification as leakage-free next-day forecasting and compared four SVM variants at five monitoring stations in Jakarta for 2021–2025. FSVM-CIL achieved the highest mean macro-F1 (0.521) and G-mean (0.502), 0.075 and 0.327 points higher than the standard SVM, and raised the pooled recall of Unhealthy$^+$ days from 0.18 to 0.44. The Friedman test did not show significant differences among the models ($\chi^2 = 5.67$, $p = 0.129$), and the best model depended on the local imbalance pattern: FSVM-CIL was best at three of the five stations, and no strategy improved on the standard SVM at the most imbalanced station. Imbalance handling should therefore be a routine part of ISPU category forecasting, with the strategy chosen for each station. Future work should use numerical weather forecasts as predictors, exploit the ordinal nature of ISPU categories, and include calibrated low-cost sensor networks.

References

Ahmar, A. S., Rais, Z., & Tunnas, F. (2025). Comparative analysis of support vector regression (SVR) and SutteARIMA in predicting coal prices in Indonesia. Quality & Quantity, 59(5), 4185–4200. https://doi.org/10.1007/s11135-025-02162-2

Annas, S., Uca, U., Irwan, I., Safei, R. H., & Rais, Z. (2022). Using k-means and self organizing maps in clustering air pollution distribution in Makassar City, Indonesia. Jambura Journal of Mathematics, 4(1), 167–176. https://doi.org/10.34312/jjom.v4i1.11883

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 6(1), 20–29. https://doi.org/10.1145/1007730.1007735

Batuwita, R., & Palade, V. (2010). FSVM-CIL: Fuzzy support vector machines for class imbalance learning. IEEE Transactions on Fuzzy Systems, 18(3), 558–571. https://doi.org/10.1109/TFUZZ.2010.2042721

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

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

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21, 6. https://doi.org/10.1186/s12864-019-6413-7

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1–30. https://jmlr.org/papers/v7/demsar06a.html

Dinas Lingkungan Hidup Provinsi DKI Jakarta. (2021). Laporan kualitas udara Jakarta, Selasa, 30 November 2021 [Jakarta air quality report, 30 November 2021]. https://lingkunganhidup.jakarta.go.id/files/73948-2021-11-30-09-08-30.pdf

Dinas Lingkungan Hidup Provinsi DKI Jakarta. (2023a). Indeks Standar Pencemaran Udara (ISPU) tahun 2021 [Data set]. Satu Data Jakarta. Retrieved November 8, 2026, from https://satudata.jakarta.go.id/open-data/detail?kategori=dataset&page_url=indeks-standar-pencemaran-udara-ispu-tahun-2021&data_no=1

Dinas Lingkungan Hidup Provinsi DKI Jakarta. (2023b). Data Indeks Standar Pencemar Udara (ISPU) di Provinsi DKI Jakarta 2023 [Data set]. Satu Data Jakarta. Retrieved November 8, 2026, from https://satudata.jakarta.go.id/open-data/detail?kategori=dataset&page_url=data-indeks-standar-pencemar-udara-ispu-di-provinsi-dki-jakarta-2023&data_no=1

Dinas Lingkungan Hidup Provinsi DKI Jakarta. (2023c). Indeks Standar Pencemaran Udara (ISPU) tahun 2022 [Data set]. Satu Data Jakarta. Retrieved November 8, 2026, from https://satudata.jakarta.go.id/open-data/detail?kategori=dataset&page_url=indeks-standar-pencemaran-udara-ispu-tahun-2022&data_no=1

Dinas Lingkungan Hidup Provinsi DKI Jakarta. (2026). Data Indeks Standar Pencemar Udara (ISPU) di Provinsi DKI Jakarta [Data set]. Satu Data Jakarta. Retrieved November 8, 2026, from https://satudata.jakarta.go.id/open-data/detail?kategori=dataset&page_url=data-indeks-standar-pencemar-udara-ispu-di-provinsi-dki-jakarta&data_no=1

Fahmuddin S, M., Rais, Z., & Yuniar, E. C. (2025). Classification of hypertension using methods support vector machine genetic algorithm (SVM-GA). ARRUS Journal of Mathematics and Applied Science, 5(1), 11–16. https://doi.org/10.35877/mathscience3976

Firdaus, Z. F., & Permadi, D. A. (2024). Determination of fine particulate matter episode periods in Jakarta by incorporating meteorological conditions. Jurnal Presipitasi, 21(2), 586–597. https://doi.org/10.14710/presipitasi.v21i2.586-597

Hafid, H., Rais, Z., & Rezky, A. R. R. T. (2025). Klasifikasi curah hujan di Kota Makassar menggunakan gradient boosting machine (GBM) [Rainfall classification in Makassar City using gradient boosting machine]. VARIANSI: Journal of Statistics and Its Application on Teaching and Research, 7(2), 133–142. https://doi.org/10.35580/variansiunm386

Handhayani, T. (2023). An integrated analysis of air pollution and meteorological conditions in Jakarta. Scientific Reports, 13, 5798. https://doi.org/10.1038/s41598-023-32817-9

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hsu, C.-W., & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks, 13(2), 415–425. https://doi.org/10.1109/72.991427

Jayadi, B. V., Lauro, M. D., Rusdi, Z., & Handhayani, T. (2024). Air quality index classification for imbalanced data using machine learning approach. SISTEMASI: Jurnal Sistem Informasi, 13(3), 951–958. https://doi.org/10.32520/stmsi.v13i3.3503

Lin, C.-F., & Wang, S.-D. (2002). Fuzzy support vector machines. IEEE Transactions on Neural Networks, 13(2), 464–471. https://doi.org/10.1109/72.991432

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Ministry of Environment and Forestry. (2020). Peraturan Menteri Lingkungan Hidup dan Kehutanan Nomor P.14/MENLHK/SETJEN/KUM.1/7/2020 tentang Indeks Standar Pencemar Udara [Regulation of the Minister of Environment and Forestry No. P.14/2020 on the Air Pollution Standard Index]. https://peraturan.go.id/files/bn774-2020.pdf

NASA Langley Research Center. (2026). Prediction of Worldwide Energy Resources (POWER) project: Daily point API, version 2 [Data set]. Retrieved November 10, 2026, from https://power.larc.nasa.gov/

Muttaqim, M. H. Z., Ruliana, R., & Rais, Z. (2023). Application of K-medoids algorithm in provincial grouping in Indonesia based on case of environmental pollution. SAINSMAT: Journal of Applied Sciences, Mathematics, and Its Education, 12(1), 30–39. https://doi.org/10.35877/sainsmat1775

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Rahmat, R. W., Annas, S., & Rais, Z. (2023). Analisis support vector regression (SVR) untuk meramalkan indeks kualitas udara di Kota Makassar [Support vector regression analysis for forecasting the air quality index in Makassar City]. VARIANSI: Journal of Statistics and Its Application on Teaching and Research, 5(3), 104–117. https://doi.org/10.35580/variansiunm107

Rais, Z., Fahmuddin S, M., Saida, S., & Triutomo, A. (2025). Implementation of machine learning algorithm with extreme gradient boosting (XGBoost) method in hypertension level classification. Journal of Applied Science, Engineering, Technology, and Education, 7(1), 126–136. https://doi.org/10.35877/454RI.asci4191

Rais, Z., Hafid, H., & Bunga, Y. R. (2026). Rainfall classification using output statistics models based on classification and regression trees with principal component analysis preprocessing. JINAV: Journal of Information and Visualization, 7(1), 69–82. https://doi.org/10.35877/454RI.jinav4788

Ridho, I. I., & Mahalisa, G. (2023). Analisis klasifikasi dataset Indeks Standar Pencemaran Udara (ISPU) di masa pandemi menggunakan algoritma support vector machine (SVM). Technologia: Jurnal Ilmiah, 14(1), 38–41. https://doi.org/10.31602/tji.v14i1.8005

Ruliana, R., Maru, R., Rais, Z., & Ahmar, A. S. (2025). Integration of SEM and nonparametric spline in spatial data modeling and visualization for analysis of CO2 reduction using green space in Makassar City. JOIV: International Journal on Informatics Visualization, 9(3), 1319–1329. https://doi.org/10.62527/joiv.9.3.3807

Ruliana, R., Rais, Z., & Sudirman, L. M. R. (2024). The support vector machine (SVM) and random forest methods for classification graduation rate. ARRUS Journal of Engineering and Technology, 4(2), 203–210. https://doi.org/10.35877/jetech3436

Salsabila, D., & Utami, N. F. (2025). Klasifikasi dan pemetaan spasial kualitas udara berbasis ISPU menggunakan support vector machine dengan data multisumber. Jurnal Sains dan Teknologi Lingkungan, 17(2), 167–187.

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

Syuhada, G., Akbar, A., Hardiawan, D., Pun, V., Darmawan, A., Heryati, S. H. A., Siregar, A. Y. M., Kusuma, R. R., Driejana, R., Ingole, V., Kass, D., & Mehta, S. (2023). Impacts of air pollution on health and cost of illness in Jakarta, Indonesia. International Journal of Environmental Research and Public Health, 20(4), 2916. https://doi.org/10.3390/ijerph20042916

Veropoulos, K., Campbell, C., & Cristianini, N. (1999). Controlling the sensitivity of support vector machines. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI '99) Workshop on Support Vector Machines (pp. 55–60).

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.