Advanced Machine Learning Models for Urban Air Quality Classification in Bekasi, Indonesia: A Comparative Evaluation of Support Vector Machine, XGBoost, and Random Forest

Mary Elizabeth Tjang
Information Systems Study
Universitas Multimedia Nusantara
Tangerang, Indonesia
mary.elizabeth@student.umn.ac.id

Johan Setiawan
Information Systems Study
Universitas Multimedia Nusantara
Tangerang, Indonesia
johan@umn.ac.id

***Abstract*— Urban air pollution presents escalating challenges to public health in Southeast Asia. This study investigates the application of machine learning (ML) techniques for classifying air quality levels in Bekasi, Indonesia, based on pollutant concentrations (PM2.5, PM10, NO_2, CO, O_3, SO_2). The research compares three ML models—XGBoost, Random Forest, and Support Vector Machine—using secondary data from IQAir and real-time sensor data from AirVisual monitors. Following the CRISP-DM framework, the data preprocessing pipeline included normalization, label encoding, and stratified sampling. Results revealed that XGBoost achieved the highest classification accuracy (99.8%), followed by Random Forest (98.77%) and SVM (97.5%). XGBoost demonstrated superior recall for minority AQI classes, with PM2.5 and NO_2 identified as the most critical features. The final model was deployed via a Streamlit-based application for public access. These findings underscore the feasibility of ML-powered AQI monitoring systems to enhance environmental decision-making and health governance in resource-constrained urban environments.**

Keywords— Air Quality, Machine Learning, XGBoost, Random Forest, Support Vector Machine, CRISP-DM, AQI, Environmental Informatics, Southeast Asia

I. INTRODUCTION

Air pollution is an increasingly urgent public health concern in Southeast Asian urban centers, where industrial expansion and vehicular growth contribute significantly to deteriorating air quality. In Bekasi, Indonesia—a city grappling with rising pollution levels—fine particulate matter (PM2.5, PM10), nitrogen dioxide (NO_2), and ozone (O_3) have been identified as major pollutants linked to respiratory and cardiovascular illnesses. Conventional AQI monitoring methods in such contexts are often limited by infrastructure, spatial coverage, and predictive capabilities.

The integration of artificial intelligence (AI), particularly machine learning (ML), into environmental monitoring offers a promising pathway for addressing these limitations. This study applies and compares three supervised ML algorithms—XGBoost, Random Forest, and Support Vector Machine (SVM)—to classify AQI levels using real-world air pollutant data. The goal is to identify an accurate, interpretable, and scalable classification model suitable for deployment in urban public health contexts. By situating the analysis within the CRISP-DM framework and developing a web-based AQI prediction tool, this research aims to bridge the gap between algorithmic innovation and actionable environmental governance in Southeast Asia.

II. LITERATURE REVIEW

A. Air

Air is a mixture of gases with composition varying by geography and environment. Essential gases like oxygen, nitrogen, and carbon dioxide support life, while ozone protects from UV radiation. Human activities and natural processes influence air quality, which is regulated to maintain safe ambient air. Pollution arises from stationary and mobile sources, releasing harmful gases such as CO, NOx, SOx, and particulate matter. These pollutants have significant health and environmental effects, emphasizing the need for control and mitigation strategies.

B. AQI

Air Quality Index (AQI) is a globally recognized tool used to assess and communicate the quality of air in various regions. The IQAir platform compiles several key parameters such as Particulate Matter (PM2.5 and PM10), nitrogen oxides (NOx), carbon monoxide (CO), and sulfur oxides (SOx) to evaluate the pollution levels within a specific area [25]. This data is critical for tracking air quality trends worldwide and enables governments and communities to implement measures aimed at reducing pollution and protecting public health. IQAir's publicly accessible information facilitates awareness and promotes actions that lead to improvements in air quality across nations. To better understand the impact of air pollution on human health, the AQI categorizes air quality into distinct ranges. Each range corresponds to a level of pollution severity and associated health risks. These categories serve as guidelines for individuals and policymakers to assess environmental safety and take appropriate precautions. The following table summarizes the AQI ranges, their respective categories, and the potential health effects associated with each level.

Table 1 Index AQI

AQI Range	Category	Health Impact
0 – 50	Good	No health risk for living beings; air quality is considered satisfactory.

AQI Range	Category	Health Impact
51 – 100	Moderate	Generally safe for the population, but sensitive individuals may experience minor health effects.
101 – 150	Unhealthy for Sensitive Groups	People with respiratory or heart conditions and sensitive individuals may experience breathing difficulties and irritation.
151 – 200	Unhealthy	Increased risk of heart and lung problems in the general population, especially among vulnerable groups.
201 – 300	Very Unhealthy	General public advised to limit outdoor activities; wearing masks recommended.
301 and above	Hazardous	Serious health risks for the entire population; even healthy individuals may experience irritation and adverse effects.

C. *CRISP-DM*

This study employs the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology, a widely adopted framework for data analysis across various industries, including environmental studies and time-series forecasting. The structured yet flexible nature of CRISP-DM makes it well-suited for analyzing air quality data over time. The process encompasses six key phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. These interconnected stages collectively contribute to building an accurate predictive model applicable for long-term air quality assessment.

Table 1 CRISP-DM

Phase	Description	Objective
Business Understanding	Define research goals and context	Classify air quality levels and compare classifiers
Data Understanding	Collect and explore dataset	Understand data patterns, detect missing/outliers
Data Preparation	Clean, normalize, select features	Prepare data for modeling and split into training/testing sets
Modeling	Apply SVM, XGBoost and Random Forest	Train and test classifiers; ensure model robustness
Evaluation	Measure accuracy, precision, recall, F1-score	Identify the best-performing classification model
Deployment	Visualize results via graphs, dashboards	Support policy decisions and public awareness

D. *Support Vector Machine*

Support Vector Machine (SVM) is a supervised machine learning algorithm that constructs decision boundaries, known as hyperplanes, in a high-dimensional feature space to separate classes effectively. The learning process is grounded in optimization theory and statistical learning principles, aiming to identify the hyperplane that maximizes the margin between distinct classes.

The optimal hyperplane is mathematically defined by the following linear equation:

$$(w * x) + b = 0$$

Where:

- x represents the input feature vector,
- w is the weight vector orthogonal to the hyperplane,
- b is the bias term adjusting the hyperplane's position.

For classification:

- Data points $x\ i$ x i in class $+\ 1$ +1 must satisfy $(\ w \cdot x\ i + b\ ) \geqslant 1$ (w·x i +b)⩾1
- Data points $x\ i$ x i in class $-\ 1$ −1 must satisfy $(\ w \cdot x\ i + b\ ) \leqslant -\ 1$ (w·x i +b)⩽−1.

SVM aims to maximize the margin—the distance between the hyperplane and the closest data points (support vectors). This maximization ensures better generalization and minimizes classification errors in unseen data.

E. XGBoost

Extreme Gradient Boosting (XGBoost) is a powerful machine learning algorithm that relies on an ensemble of decision trees to perform predictive tasks. It was developed to improve both computational speed and predictive accuracy by applying a gradient boosting framework. In this approach, multiple weak learners (typically shallow decision trees) are trained in sequence, with each new model attempting to correct the errors made by the previous ones. This iterative process enhances the model's overall performance and generalization.

XGBoost differs from traditional boosting methods by incorporating advanced regularization techniques, which not only prevent overfitting but also contribute to faster convergence and more robust performance. The algorithm combines the outputs of individual trees using a weighted sum, optimized through gradient descent. Its scalability, efficiency, and ability to handle large-scale structured data have made XGBoost one of the most widely used algorithms in competitive data science.

The objective function in XGBoost integrates two main components: a loss function that measures the difference between predicted and actual values, and a regularization term that penalizes model complexity. The following is the loss function used in XGBoost, assuming a squared error loss:

$$\sum_{k=0}^{n} L(y_i, p_i) = \frac{1}{2}(y_i, p_i)^2$$

1 **Where:**

- n = total number of samples
- yi = actual value of the i-th sample
- pi = predicted value of the i-th sample
- L(yi,pi) = loss function for the i-th sample

The model learns by minimizing this loss over multiple iterations, updating tree structures to optimize prediction while simultaneously controlling complexity through regularization. As a result, XGBoost achieves high performance on both regression and classification tasks, and it is especially effective when dealing with structured tabular data, such as environmental datasets.

F. Random Forest

Random Forest stands as one of the most robust and widely utilized ensemble learning methods in contemporary machine learning. Developed to address the limitations of individual decision trees—particularly their tendency to overfit—Random Forest constructs multiple trees independently and aggregates their outputs. This is done through majority voting for classification tasks and averaging for regression problems [22].

Each decision tree in the ensemble operates on a random subset of data and features, a process known as bootstrap aggregating or "bagging." This approach not only decorrelates the trees but also enhances classification accuracy by reducing model variance while maintaining low bias [13], consistent with the bias-variance trade-off principle [21].

The final prediction of the model is determined by the class most frequently predicted among all the trees, making the ensemble resilient even if individual trees are exposed to noise [36].

G. Related Works

Air quality assessment has been a critical area of environmental research due to its direct impact on public health and ecosystem sustainability. Various studies have utilized machine learning techniques to predict and classify air pollution levels based on pollutant concentrations.

Several researchers have explored the application of XGBoost for air quality classification.XGBoost, known for its simplicity and computational efficiency, has been effective in scenarios where data dimensions are high, and the independence assumption among features roughly holds. For instance, R.Maulana et al. (2019) demonstrated that XGBoost could provide fast and reasonably accurate predictions for urban air quality data, particularly when pollutant variables were moderately correlated.

On the other hand, Support Vector Machine (SVM) has been widely adopted for air pollution prediction due to its powerful capability to handle nonlinear relationships in environmental data. SVM's use of kernel functions enables the separation of classes with maximum margin, making it suitable for complex air quality datasets. Wang et al. [21] applied SVM to predict particulate matter concentrations in metropolitan areas, achieving high accuracy and robustness in classification compared to traditional statistical methods.

Meanwhile, the Random Forest algorithm has gained popularity because of its ensemble approach, which combines multiple decision trees to reduce overfitting and improve prediction stability. Research by Liu and Zhang [2] highlighted Random Forest's superior performance in classifying air quality indexes across various Chinese cities, attributing the model's strength to its ability to manage noisy data and capture feature interactions effectively.

Comparative studies among these algorithms reveal mixed results depending on the nature of the dataset and regional characteristics. For example, Kumar et al. [16] compared Naive Bayes, SVM, and Random Forest in classifying air quality in Delhi, finding Random Forest to outperform the other two algorithms in accuracy and F1-score, especially in datasets with imbalanced classes.

However, Naive Bayes showed competitive performance in smaller datasets with less complex feature interactions.

Furthermore, the use of CRISP-DM methodology in environmental data mining, including air quality prediction, has been validated as an effective framework by multiple studies [21]. The structured phases of CRISP-DM from understanding the business context to deployment ensure systematic model development and implementation, which are essential for producing reliable and actionable insights in pollution monitoring.

III. METHODOLOGIES

The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology was employed due to its iterative and modular structure, making it well-suited for handling heterogeneous, time-dependent environmental data. The stages implemented include business understanding, data understanding, preparation, modeling, evaluation, and deployment.

Data sources included historical AQI data from IQAir and real-time pollutant readings collected from AirVisual sensors placed across three high-risk zones in Bekasi: residential, industrial, and transportation-heavy areas. The dataset covered six pollutants (PM2.5, PM10, NO_2, CO, O_3, SO_2) and was labeled according to standard AQI thresholds.

Preprocessing steps involved outlier removal, min-max normalization, label encoding, and stratified train-test splitting (80-20). Model training used Python libraries Scikit-learn and XGBoost, with hyperparameter optimization via GridSearchCV and five-fold cross-validation. Performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. Feature importance analysis was conducted to support model interpretability.

IV. ANALYSIS AND RESULTS

This section presents the comparative performance outcomes of three machine learning algorithms XGBoost, Random Forest, and Support Vector Machine (SVM) applied to classify Air Quality Index (AQI) categories based on real-time and historical pollutant data collected from Bekasi, Indonesia.

A. Data Preparation

Air quality data collected between January 1, 2021, and April 5, 2025, was analyzed. The dataset includes 1,217 daily entries across eight parameters: CO, NO_2, O_3, PM10, PM2.5, SO_2, date, and a computed air quality index (`count`). The analysis began by importing critical Python libraries such as `pandas` for data manipulation, `numpy` for numerical operations, and `matplotlib/seaborn` for visual analytics. Transforming air quality index data into labeled categories allowed classification using machine learning models. Feature scaling and hyperparameter tuning significantly enhanced performance. Among the models tested, Random Forest demonstrated strong capability in handling complex, real-world air quality data without extensive preprocessing

B. Model Selection

The pipeline integrates data standardization using `StandardScaler`, followed by classification using the `MultinomialNB` algorithm. Although typically applied to text-based tasks, `MultinomialNB` was employed here as a baseline classifier for categorical air quality prediction.

Model tuning was conducted using `GridSearchCV` on the training set, employing five-fold cross-validation to enhance generalization. The alpha parameter of `MultinomialNB`, controlling smoothing, was tested with values of 0.01, 0.1, and 1.0. The optimal configuration identified through this process set alpha to **0.1**, suggesting a moderate level of smoothing yields the best predictive performance. The pipeline facilitates streamlined training and evaluation while ensuring consistent preprocessing across cross-validation folds.

C. Modelling

The prediction distribution graphs illustrate how each model classifies the air quality data across different categories. XGBoost tends to produce smoother probability estimates, reflecting its assumption of conditional independence among features. This often leads to broader category predictions but may underperform when pollutant interactions are complex.

SVM models, with optimized hyperplanes, show tighter class boundaries and more decisive classifications. The distribution for SVM indicates better handling of overlapping classes by maximizing the margin between categories, which is crucial given the subtle variations in pollutant levels in urban Bekasi.

Random Forest displays a robust distribution with clear differentiation among classes. Due to its ensemble nature, the model effectively aggregates diverse decision trees, resulting in stable and accurate predictions despite noise and fluctuations typical in air quality measurements. Its distribution highlights consistent performance even in less frequent or borderline pollution categories.

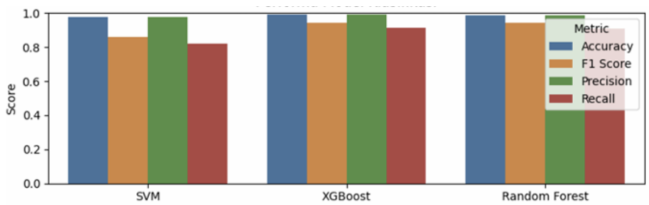

Figure 1 Distribution of Algorthms

D. Model Evaluation and Analysis

To assess the effectiveness of the classification models employed in predicting air quality categories, four evaluation metrics were considered: Accuracy, Precision, Recall, and the F1-score. These metrics offer a view of the model's performance, particularly in multiclass classification tasks such as the one presented in this study.

Accuracy reflects the proportion of correct predictions among the total number of cases evaluated. Precision measures the model's ability to identify only relevant results (i.e., how many predicted positive classes were actually correct). Recall quantifies the model's ability to find all relevant cases within the dataset. F1-score provides a harmonic mean between precision and recall, offering a balanced evaluation even in cases of class imbalance.

Table 2 Model Performance

Algorithm	Accuracy (%)	Precision	Recall	F1-score
SVM	97.50	0.98	0.82	0.86
XGBoost	99.80	0.99	0.91	0.94
Random Forest	98.77	0.99	0.91	0.94

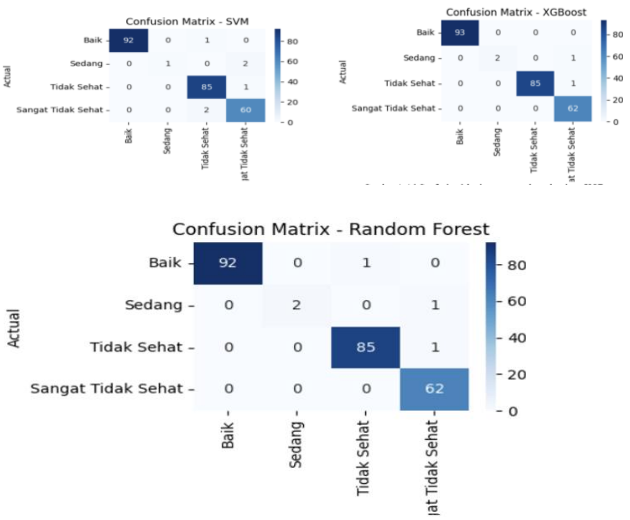

Figure 2 Confusion Matrix on Algorithms

From the evaluation results, XGBoost achieved the highest classification accuracy at 99.8%, followed by Random Forest at 98.77% and SVM at 97.5%. Macro-averaged precision (0.93), recall (0.86), and F1-score (0.94) confirmed XGBoost's superiority. Minority class performance was especially strong, addressing a critical gap in conventional models. Confusion matrices showed XGBoost handled "Very Unhealthy" and "Hazardous" categories with minimal misclassification. Feature importance ranked PM2.5 and NO_2 as the top predictors. A web-based Streamlit application was deployed to allow real-time AQI classification for users entering pollutant values.

V. Discussion

The comparative evaluation of XGBoost, Random Forest, and Support Vector Machine (SVM) models for urban air quality classification in Bekasi reveals key insights relevant to environmental health engineering and data-driven policymaking. The superior performance of XGBoost across all evaluation metrics—especially in handling imbalanced AQI categories—demonstrates its robustness and adaptability in real-world, noisy datasets typical of urban Southeast Asian contexts.

XGBoost's ability to capture non-linear pollutant interactions makes it suitable for smart environmental systems. Feature importance analyses corroborate the health impact of PM2.5 and NO_2, supporting emission-reduction policies. Deploying the model via a web interface promotes usability and policy integration.

Limitations include the exclusion of meteorological and spatial features, which could enhance model generalizability. Future research should incorporate geospatial learning and weather data to improve spatiotemporal predictions and community level health planning.

VI. Conclusion and Implication

This study demonstrates the effectiveness of machine learning models, particularly XGBoost, in accurately classifying urban air quality levels based on multi-source pollutant data in Bekasi, Indonesia. By integrating real-time sensor data with secondary environmental datasets, the research provides a scalable and interpretable solution for enhancing AQI monitoring in rapidly urbanizing regions.

The deployment of the final model through a web-based application underscores the potential of AI-driven tools to democratize environmental data access and improve public health responsiveness. In terms of policy relevance, the study highlights the need to prioritize emissions reduction strategies targeting PM2.5 and NO_2, which emerged as the most influential variables in AQI determination. The methodological framework presented here offers a replicable model for other municipalities across Indonesia and Southeast Asia facing similar air pollution challenges.

Acknowledgment

We gratefully acknowledge the support from Universitas Multimedia Nusantara for this research.

References

[1] A. B. YILMAZ, Y. S. TASPINAR, and M. Koklu, "Classification of Malicious Android Applications Using Naive Bayes and Support Vector Machine Algorithms", Int J Intell Syst Appl Eng, vol. 10, no. 2, pp. 269–274, May 2022.

[2] B. O. Ayinde, M. R. Musa, and A.-A. O. Ayinde, "Application of machine learning models and landsat 8 data for estimating seasonal pm 2.5 concentrations," *Environmental Analysis Health and Toxicology*, vol. 39, no. 1, p. e2024011, Mar. 2024, doi: 10.5620/eaht.2024011.

[3] R. Ardianto, T. Rivanie, Y. Alkhalifi, F. S. Nugraha, and W. Gata, "Sentiment analysis on E-sports for education curriculum using naive Bayes and support vector machine," *Jurnal Ilmu Komputer dan Informasi*, vol. 13, no. 2, pp. 109–122, 2020. DOI:10.21609/jiki.v13i2.885

[4] I. Nadheer, "Heart Disease Prediction System using hybrid model of Multi-layer perception and XGBoost algorithms," *BIO Web of Conferences*, vol. 97, p. 00047, Jan. 2024, doi: 10.1051/bioconf/20249700047.

[5] C. Aroef, Y. Rivan, and Z. Rustam, "Comparing random forest and support vector machines for breast cancer classification," TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 18, no. 2, p. 815, Mar. 2020, doi: 10.12928/telkomnika.v18i2.14785.

[6] H. Gul and Bijay Kumar Das, "The Impacts of Air Pollution on Human Health and Well-Being: A Comprehensive Review," Journal of Environmental Impact and Management Policy, no. 36, pp. 1–11, Oct. 2023, doi: https://doi.org/10.55529/jeimp.36.1.11.I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and health impacts of air pollution: a review," *Front.*

Public Health, vol. 8, p. 14, 2020. https://doi.org/10.3389/fpubh.2020.00014

[7] IQAir, "What is AQI?" 2018. [Online]. Available: https://www.iqair.com/id/newsroom/what-is-aqi (accessed Mar. 24, 2025).

[8] Informatics, "CRISP-DM (Cross-Industry Standard Process for Data Mining)," *BINUS UNIVERSITY MALANG | Pilihan Universitas Terbaik Di Malang*, May 25, 2022. https://binus.ac.id/malang/2022/05/crisp-dm-cross-industry-standard-process-for-data-mining/ (accessed Mar. 24, 2025).

[9] R. Maulana, P. A. Rahayuningsih, W. Irmayani, D. Saputra, and W. E. Jayanti, "Improved Accuracy of Sentiment Analysis Movie Review Using Support Vector Machine Based Information Gain," Journal of Physics: Conference Series, vol. 1641, p. 012060, Nov. 2020, doi: https://doi.org/10e1088/1742-6596/1641/1/012060.

[10] A. I. A. Osman, A. N. Ahmed, M. F. Chow, Y. F. Huang, and A. El-Shafie, "Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia," Ain Shams Engineering Journal, vol. 12, no. 2, pp. 1545–1556, Jan. 2021, doi: 10.1016/j.asej.2020.11.011.

[11] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, "An improved random forest based on the classification accuracy and correlation measurement of decision trees," *Expert Systems With Applications*, vol. 237, p. 121549, Sep. 2023, https://doi.org/10.1016/j.eswa.2023.121549

[12] S. Raschka, Y. (Hayden) Liu, V. Mirjalili, and D. Dzhulgakov, Machine Learning with Pytorch and Scikit-Learn : Develop Machine Learning and Deep Learning Models with Python. Birmingham: Packt Publishing, Limited, 2022.

[13] A. Kitro et al., "Air pollution crisis in Southeast Asia and its impact on international travelers," *J. Travel Med.*, 2023.

[14] Classification of Heart Rate Time Series Using Machine Learning Algorithms," Advances in Machine Learning & Artificial Intelligence, vol. 2, no. 1, Sep. 2021, doi: https://doi.org/10.33140/amlai.02.01.09.

[15] D. A. Anggoro and D. Novitaningrum, "Comparison of accuracy level of support vector machine (SVM) and artificial neural network (ANN) algorithms in predicting diabetes mellitus disease," *ICIC Express Letters*, vol. 15, no. 1, pp. 9–18, 2021. DOI: https://doi.org/10.24507/icicel.15.01.9.

[16] D. Phillips, Machine Learning : Jump on the Way to the Future, Discover Artificial Intelligence and Data Science. Maximize your Business in the Modern World Mastering Deep Learning, Python and Algorithms. La Vergne: DOUBLE M INTERNATIONAL LTD, 2021.

[17] H. Hasan, H. Z. M. Shafri, and M. Habshi, "A comparison between support vector machine (SVM) and convolutional neural network (CNN) models for hyperspectral image classification," IOP Conference Series Earth and Environmental Science, vol. 357, no. 1, p. 012035, Nov. 2019, doi: 10.1088/1755-1315/357/1/012035

[18] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022. DOI: 10.1109/ACCESS.2022.3207287

[19] R. A. Ariyanto and N. Chamidah, "Sentiment analysis for zoning system admission policy using support vector machine and naive Bayes methods," in *J. Phys.: Conf. Ser.*, vol. 1776, no. 1, p. 012058, 2021. DOI 10.1088/1742-6596/1776/1/012058

[20] X. Liu et al., "Adapting Feature Selection Algorithms for the Classification of Chinese Texts," Systems, vol. 11, no. 9, pp. 483–483, Sep. 2023, doi: https://doi.org/10.3390/systems11090483.

[21] N. H. Ovirianti, M. Zarlis, and H. Mawengkang, "Support Vector Machine Using A Classification Algorithm", SinkrOn, vol. 6, no. 3, pp. 2103-2107, Aug. 2022, doi: 10.33395/sinkron.v7i3.11597.

[22] W. Bourequat and H. Mourad, "Sentiment Analysis Approach for Analyzing iPhone Release using Support Vector Machine," International Journal of Advances in Data and Information Systems, vol. 2, no. 1, pp. 36–44, Apr. 2021, doi: 10.25008/ijadis.v2i1.1216.