Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 8, No. 3, November 2025, pp. 677 – 686
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Comparative Evaluation of Optuna-Optimized Radial Basis Function and Sigmoid Kernels in Support Vector Machine for Smart Air Quality Classification

[1]Nanda Galea, [2]A. Rahman, [3*]Renny Maulidda, [4]Nyayu Latifah Husni
[1,2,3,4]Departement of Applied Electrical Engineering, Politeknik Negeri Sriwijaya, Indonesia
Email: [1]nanndagaleaa07@gmail.com, [2]a.rahman.600@yahoo.com,
[3]rennymaulidda@polsri.ac.id, [4]nyayu_latifah@polsri.ac.id

Article Info

Article history:
Received Jun 21th, 2025
Revised Nov 14th, 2025
Accepted Nov 25th, 2025

Keyword:
Air Quality Classification
Optuna
Radial Basis Function
Sigmoid
Support Vector Machine

ABSTRACT

Poor air quality can have a serious impact on human health, so a classification system capable of accurately identifying air conditions is needed. This research proposes an air quality classification method using the Support Vector Machine (SVM) algorithm with two types of non-linear kernels, namely Radial Basis Function (RBF) and Sigmoid. The data used is obtained from various environmental sensors that record parameters such as CO, smoke, HC, TVOC, eCO_2, temperature, and humidity, and then collected in the form of historical datasets. To enhance the accuracy and efficiency of the model, hyperparameter optimization was performed automatically using Optuna. The evaluation results showed that the SVM with an RBF kernel performed better than the SVM with a sigmoid kernel, achieving an accuracy value of 96.67% and an F1-score of 96.80%. Additionally, RBF demonstrated higher stability in a 5-fold cross-validation. This research demonstrates that the combination of SVM and Optuna is effective in developing an accurate air quality classification system, with potential for further development as a sensor-based air monitoring system and Internet of Things (IoT) solution.

Copyright © 2025 Puzzle Research Data Technology

Corresponding Author:
Renny Maulidda,
Departement of Applied Electrical Engineering,
Politeknik Negeri Sriwijaya,
Srijaya Negara Road, Bukit Besar 30139, Palembang City, Indonesia.
Email: rennymaulidda@polsri.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v8i3.37563

1. INTRODUCTION

Poor air quality has become an increasingly urgent global concern, impacting not only human health but also the environment and the overall quality of life in urban areas. According to the World Health Organization (WHO), more than 91% of the global population is exposed to unhealthy air, and the State of Global Air 2019 report by the Health Effects Institute (HEI) identifies air pollution as the fifth leading cause of death worldwide, accounting for nearly five million premature deaths in 2017 [1]. Continuous exposure over a long duration to pollutants can lead to various diseases, including chronic inflammation of the respiratory tract, lung infections, clogged arteries, and plaque formation in the blood vessels of the brain [2].

In Indonesia, particularly in Jakarta, air pollution has become a serious environmental and public health problem. A study by Anwar et al. (2023) revealed that air pollution in Jakarta resulted in over 10,000 premature deaths and more than 5,000 hospital admissions annually, imposing an economic burden of USD 2.94 billion (2.2% of the city's Gross Regional Domestic Product (GRDP)) [3]. This alarming impact underscores the urgent need for an accurate and intelligent air quality classification system to support environmental policy and public health protection. Given the extensive impact on public health, efforts to classify air quality based on environmental data are becoming increasingly relevant, especially with the help of technology. Various sensing devices can now be used to record environmental conditions more accurately, and the data obtained from these measurements can be organised into structured datasets for research purposes. In processing and classifying the data obtained, an effective and highly accurate multidimensional classification method such as Support Vector Machine (SVM) is required [4], [5], [6].

SVM is a widely utilized machine learning algorithm for performing classification tasks. SVM functions by transforming input data into a feature space and identifying the most suitable hyperplane to divide different classes [7]. In the classification stage, the kernel transforms the data into a high-dimensional feature space. The constructed hyperplane distinguishes the classes by maximizing the gap between itself and the closest points from each class, known as support vectors [8], [9]. Choosing the right kernel in SVM has a significant impact on the model's performance. Two non-linear kernels that are often used are Radial Basis Function (RBF) and Sigmoid. Each has characteristics in mapping data to different feature spaces, so it can produce classification results that vary depending on the characteristics of the data being analysed.

In addition to kernel selection, hyperparameter tuning is necessary for optimal SVM model performance. One of the frameworks for automatically performing optimization is Optuna. Recent studies have confirmed that Optuna significantly improves model performance and computational efficiency compared to traditional optimization techniques. For instance, in rainfall prediction using Long Short-Term Memory (LSTM) networks, Optuna achieved higher accuracy and lower computational cost than Grid Search, demonstrating its capability to identify optimal hyperparameter configurations efficiently [10]. Exploration of the hyperparameter space is achieved in this framework through the use of search strategies like the Tree-structured Parzen Estimator (TPE) [11]. By using this approach, the optimization process becomes more targeted and time-efficient, so as to produce the best parameter configuration for each type of kernel used [12]. To determine the extent of classification success, model performance can be evaluated using evaluation metrics like accuracy, precision, recall, and F1-score [13]. The higher the model performance, the more reliable the model is in classification [14].

Several previous studies have demonstrated that SVM is a highly reliable algorithm for classification tasks, particularly in mapping air quality data into meaningful categories. SVM has been successfully applied to classify air conditions based on parameters such as CO, CO_2, HC, PM_{10}, and temperature with accuracy rates exceeding 95% [15]. Other studies using Jakarta's open environmental data reported that the RBF kernel consistently produced dependable results in identifying air pollution patterns [16], while time-based Air Quality Index (AQI) prediction using SVM also proved effective for medium-term environmental assessment [17]. Additionally, the RBF kernel has shown high accuracy in gas classification systems for detecting substances such as ethanol, methanol, and acetone [18]. Although SVM performs well in classification tasks, its accuracy depends heavily on precise parameter tuning. Manual tuning is time-consuming and may not yield optimal configurations. Recent studies have shown that Optuna, a framework based on Bayesian optimization, can automatically determine optimal hyperparameter combinations more efficiently, enhancing model accuracy and stability while reducing computational costs [19], [20], [21], [22].

Building upon these findings, this study introduces a comparative evaluation of two non-linear SVM kernels (RBF and Sigmoid) each optimized automatically using Optuna. Unlike previous research that relied mainly on public datasets or limited pollutant parameters, this study utilizes real environmental sensor data comprising CO, smoke, HC, TVOC, eCO_2, temperature, and humidity. The classification is divided into four categories: Good, Moderate I, Moderate II, and Unhealthy. Through this approach, the study aims to identify which kernel configuration, under Optuna-based optimization, produces superior accuracy, consistency, and computational efficiency. The results are expected to provide deeper insights into how hyperparameter optimization and kernel selection influence SVM performance on complex, multidimensional environmental datasets, contributing to the development of a more intelligent and adaptive air quality classification model.

2. RESEARCH METHOD

The framework of the classification and evaluation process using SVM with Optuna optimization as presented in Figure 1.

The classification procedure commences with obtaining historical data from various sensors with the parameters CO, smoke, HC, TVOC, eCO_2, temperature and humidity. The historical data is stored in the form of a structured table or what is known as a dataset. Furthermore, the data will be preprocessed by cleaning, normalising and dividing the data. After the dataset is good, classification will be carried out using SVM RBF and Sigmoid modelling with hyperparameter tuning using Optuna to get a combination of C and gamma to optimise the classification process. Finally, the model will be evaluated with accuracy, precision, recall and F1-score for each kernel to determine which model is better.

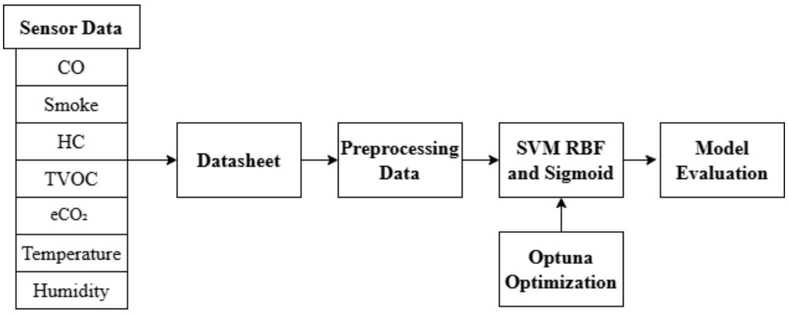

Figure 1. Classification and Evaluation Process Framework Using SVM

2.1 Data Collection Design

The dataset used in this work was obtained from various environmental sensors that record several air quality parameters. MICS 5524 sensors are used to measure CO levels, MQ-2 for smoke, TGS 2610 for HC, ENS160 for TVOC and eCO_2 and AHT21 for temperature and humidity measurements. Sensor data readings will be sent by ESP32 via the internet to Firebase. To store the data historically, it is necessary to use a spreadsheet that can record data every minute. The data collection design is illustrated in Figure 2.

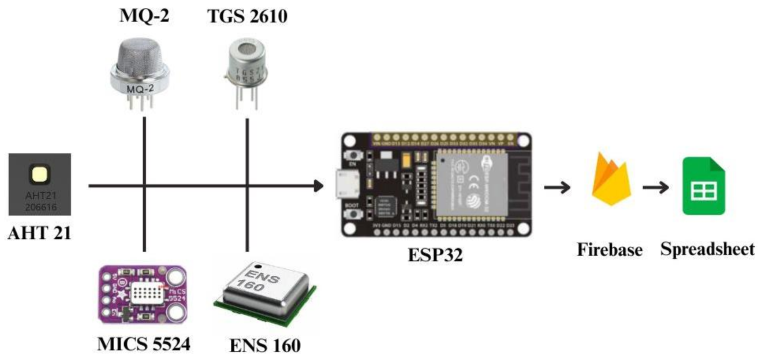

Figure 2. Data Collection Design

Each input data in this study is associated with an air quality category label divided into four classes, namely Good, Moderate I, Moderate II, and Unhealthy. The Good category indicates air conditions that are clean and free from harmful pollutants. Moderate I describe conditions with mild levels of pollution, usually from lighter-than-air substances such as carbon monoxide (CO) and smoke. Moderate II reflects the presence of moderate levels of pollution from heavier substances such as hydrocarbons (HC) and carbon dioxide equivalents (eCO_2), or from compounds that require more extensive ventilation such as total volatile organic compounds (TVOC). Meanwhile, the Unhealthy category indicates high levels of pollution that could potentially harm health if exposed for a certain amount of time.

2.2 Preprocessing Data

Data preprocessing is necessary before it is used in model training because it will affect the model performance results [23], [24]. To obtain good data quality, several preprocessing stages are carried out, namely data cleaning, normalisation, and dataset separation. The data cleaning stage is carried out to handle missing values and eliminate data considered as noise or outliers that can interfere with the training process. Next, normalisation is performed so that all features are on a comparable scale, as well as the selection of relevant features so that the model focuses more on the information that is really needed [25]. After that, the data is divided into two parts, namely training data (80%) and testing data (20%), in order to evaluate model performance fairly and objectively. This phase is conducted in a structured manner to ensure the data is fully prepared for generating an optimal classification model.

2.3 Support Vector Machine (SVM)

SVM is a classification algorithm designed to work well on complex data, by establishing the best possible separator that differentiates the data of two classes. SVMs do not only focus on accuracy on training data, but are also designed to produce models that work well on new data. This is what makes SVMs known for their strong generalisation capabilities. In order to handle data that cannot be separated directly, SVM uses a kernel function. This function helps embed the data into a higher dimensional space so that the separation pattern becomes clearer. In this study, two types of kernels are used, namely RBF and Sigmoid.

RBF kernels are excellent for data with complex non-linear patterns, while Sigmoid kernels work similarly to activation functions in neural networks. The selection of an appropriate kernel significantly impacts the classification results, as each kernel has a distinct approach to examining and separating the data [26].
The function of the RBF Kernel is:

$$K(x_i, x_j) = \exp\left(-\frac{||x-y||^2}{2\sigma^2}\right) \tag{1}$$

The function of the Sigmoid Kernel is:

$$K(x_i, x_j) = \tanh(ax_i . x_j + b) \tag{2}$$

2.4 Optuna-Optimization

Optuna is a modern innovation in hyperparameter optimization, offering a more efficient and automated method that leverages Bayesian optimization and the TPE. Unlike traditional approaches such as Grid Search and Random Search, Optuna constructs a probabilistic model to understand the relationship between hyperparameters and the objective function. This allows it to make more informed and strategic decisions about which hyperparameter values to explore next [27]. The framework is flexible as it can be used to find the maximum or minimum value of a metric, depending on the need. Additionally, Optuna supports various search methods and features a straightforward usage flow, ranging from defining functions to running optimizations. The operation of Optuna can be seen more clearly in Figure 3.

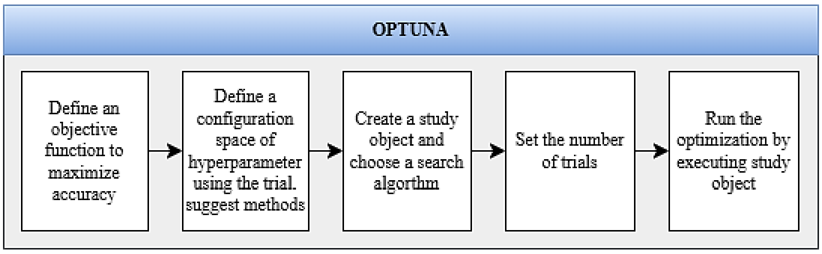

Figure 3. Optuna Workflow

Optuna begins by defining an objective function that aims to maximise model accuracy. Next, the search space for several important hyperparameters, namely C (regulation parameter), gamma (parameter for RBF kernel), coef0 (additional parameter for Sigmoid kernel), as well as the selection of kernel type between RBF or Sigmoid, is determined. After that, a study object is created to run the process of finding the best parameter combination. Tuning was performed through 50 trials on each model, and each combination was tested using 5-fold cross-validation to avoid overfitting and ensure stable results. The process ended by running an optimization to obtain the configuration that produced the best classification performance.

2.5 Model Evaluation

To assess the effectiveness of the classification model, a confusion matrix is employed. This matrix is a tabular representation that compares the model's predictions with the actual outcomes. It comprises four key elements: True Positive (TP), which refers to correctly identified positive instances; True Negative (TN), representing negative instances accurately classified as negative; False Positive (FP), which occurs when negative instances are incorrectly predicted as positive; and False Negative (FN), where positive instances are mistakenly classified as negative [28]. From these components, several evaluation matrix can be calculated, namely Accuracy, Precision, Recall and F1-Score.

Accuracy shows the proportion of total correct predictions, both positive and negative. The mathematical expression is given in Equation 3.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \tag{3}$$

Precision evaluates how accurate the positive predictions are and is determined based on Equation 4.

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\% \tag{4}$$

Recall or sensitivity indicates how well the model captures all the correct positive data. The mathematical expression is given in Equation 5.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \times 100\% \quad (5)$$

Meanwhile, F1-Score serves to harmonize precision and recall, making it especially useful when dealing with unbalanced data. The mathematical expression is given in Equation 6.

$$\text{F1 Score} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \times 100\% \quad (6)$$

2.6 Literature Review

This section reviews previous studies that applied various machine learning and optimization techniques to enhance prediction accuracy and classification performance in both health and environmental fields. These studies demonstrate how algorithms such as SVM, Random Forest, XGBoost, and other ensemble methods can effectively process complex datasets to produce reliable predictive outcomes. Optimization tools like Optuna and GridSearchCV are frequently used to tune model parameters, further improving model efficiency and accuracy. The summary of related studies is shown in Table 1.

Table 1. Literature Review

No.	Author & Year	Method Used	Research Attributes	Research Findings
1.	Tikaningsih et al. (2024) [29]	XGBoost, Random Forest, Decision Trees, CatBoost, and Extra Trees models optimized with Optuna	The research focused on predicting mortality risk and hospitalization duration for stroke patients using performance metrics such as accuracy and AUC	XGBoost achieved the best performance (mortality accuracy 86%, AUC 0.87; hospitalization accuracy 82%, AUC 0.79), showing potential to help hospitals identify high-risk patients, enhance treatment planning, and reduce costs
2.	Sholehurrohman et al. (2025) [30]	SVM algorithm with various kernels and parameter optimization using GridSearchCV	The research focused on improving air quality classification accuracy and identifying significant features such as CO concentration and proximity to industrial areas	The Polynomial Kernel achieved the best performance, with an average accuracy of 90.75%, demonstrating the SVM model's effectiveness in supporting air pollution management and mitigation efforts
3.	Ainul Yaqin et al. (2025) [31]	Random Forest algorithm with hyperparameter optimization using Optuna	The research focused on improving prediction accuracy, recall, and AUC values to enhance the effectiveness of early lung cancer identification	The optimized Random Forest model achieved 98.6% accuracy, 100% recall for the positive class, 97% for the negative class, and an AUC of 1, demonstrating exceptional performance for early lung cancer detection
4.	Gian et al. (2025) [32]	Naïve Bayes, K-NN, Random Forest, SVM, Decision Tree, AdaBoost, XGBoost, CatBoost, and LightGBM, evaluated through K-fold cross-validation	The research focused on classifying hepatitis into four categories (Acute Hepatitis, Chronic Hepatitis, Liver Abscess, and Parasitic/Viral Infections)	The SVM with a linear kernel achieved the best performance with 87% accuracy and balanced F1-scores across all classes, demonstrating the effectiveness of machine learning for early hepatitis detection

3. RESULTS AND ANALYSIS

3.1 Dataset

The dataset used in this study was collected from environmental sensors that measure several key parameters affecting air quality. These parameters include CO, smoke, HC, TVOC, estimated carbon dioxide (eCO_2), humidity, and temperature. Each record represents one data capture instance from the monitoring system. The dataset comprises 600 data entries representing various air quality conditions, categorized as Good, Moderate I, Moderate II, and Unhealthy.

Table 2. Dataset

No.	CO (ppm)	Smoke (adc)	HC (adc)	TVOC (ppb)	eCO_2 (ppm)	Humidity (%)	Temperature (˚C)	Condition
1.	0	1503	636	27	407	71	32	Good
2.	0	1499	640	33	421	71	32	Good
3.	0	1505	649	27	408	69	33	Good
4.	0	1504	656	32	418	69	33	Good

No.	CO (ppm)	Smoke (adc)	HC (adc)	TVOC (ppb)	eCO_2 (ppm)	Humidity (%)	Temperature (˚C)	Condition
5.	0	1505	655	41	438	67	33	Good
...	...	...	...	...	...	...	...	...
600.	2	1506	650	29503	17136	54	35	Unhealthy

Table 2 shows an overview of the dataset, which contains seven numerical features and one categorical target variable (Condition) that serves as the classification label. These features were later processed and used as model input for air quality classification using the SVM algorithm optimized by Optuna.

3.2 Preprocessing Data

The dataset was collected from real environmental sensors, comprising 600 records with seven numerical features (CO, Smoke, HC, TVOC, eCO_2, Humidity, and Temperature) and one categorical target (Condition), which indicates air quality levels (Good, Moderate I, Moderate II, and Unhealthy). No missing values were found, ensuring data completeness. The target variable was encoded numerically using Label Encoding, and the dataset was split into 80% training and 20% testing using stratified sampling to maintain class balance. All numerical features were normalized using StandardScaler so that each had a mean near zero and a standard deviation of one. This step ensured consistent feature scaling and prevented bias during model training. A Pipeline was used to apply preprocessing consistently, with class_weight = 'balanced' to handle class imbalance. Summary statistics of the normalized features training data can be seen in Table 3.

Table 3. Summary Statistics of Normalized Features (Training Data)

	CO	Smoke	HC	TVOC	eCO_2	Humidity	Temperature
count	480.000	480.000	480.000	480.000	480.000	480.000	480.000
mean	-0.000	-0.000	-0.000	0.000	0.000	-0.000	0.000
std	1.001	1.001	1.001	1.001	1.001	1.001	1.001
min	-0.406	-1.180	-0.435	-0.336	-0.363	-1.180	-4.048
25%	-0.406	-0.408	-0.368	-0.327	-0.344	-0.703	-0.858
50%	-0.406	-0.131	-0.311	-0.308	-0.289	-0.411	0.335
75%	-0.406	-0.024	-0.259	-0.253	-0.230	0.351	0.828
max	4.704	4.708	4.175	5.172	5.182	4.009	1.376

After normalization, all features were distributed evenly with similar scales, confirming that the preprocessing process worked correctly and that the dataset was ready for model training and optimization using Optuna. This step ensured that no single feature dominated the learning process, allowing the SVM model to achieve balanced and accurate classification results.

3.3 Optuna-Optimized Hyperparameter

Table 4 presents the hyperparameter tuning results using Optuna on two types of SVM kernels, namely RBF and Sigmoid. Optuna automatically explores the parameter search space to find the best configuration based on the highest F1-score value in the cross validation process. The optimized parameters include C as a regularization controller, gamma as a distance sensitivity parameter in the kernel, and coef0, which holds a vital role in the shape of the Sigmoid kernel function.

Table 4. Optuna-Optimize Hyperparameter

Parameter	SVM RBF Kernel	SVM Sigmoid
C	84.4454	98.7684
Gamma	0.0116	0.0045
Coef0	-	0.3058

A large value of C indicates a low degree of regularization, allowing the model to be more flexible to the training data. Gamma determines the extent to which an individual data instance contributes to the model's decision, and a small value makes the influence more diffuse. Coef0 is only used in Sigmoid kernels, serving as an additional constant in the kernel function. In the RBF kernel, this value is not used so it is marked with a (-) sign.

3.4 The Result of Model Evaluation

After the training process and hyperparameter optimization were completed, both models were tested using pre-separated test data. The evaluation is done using four main metrics: Accuracy, Precision, Recall, and F1-score, which can be seen in Table 5. These four evaluation metrics provide a comprehensive understanding of the model's classification performance, encompassing both its accuracy and its responsiveness to various data classes.

Table 5. Performance of the Model

Model SVM	Accuracy	Precision	Recall	F1-Score
RBF Kernel	96.67%	97.22%	96.67%	96.80%
Sigmoid Kernel	95.83%	96.45%	95.83%	95.96%

Table 5 shows that both models have high classification performance. In general, the model with the RBF kernel performs slightly better. With an accuracy of 96.67%, this model accurately recognizes the data pattern. The high precision and recall, at 97.22% and 96.67%, respectively, demonstrate that the model not only predicts classes accurately but also effectively recognizes all relevant data. The balanced F1-score value of 96.80% reinforces the model's stability in handling diverse data. Meanwhile, the model with the Sigmoid kernel also showed good results, with 95.83% accuracy, 96.45% precision, 95.83% recall, and 95.96% F1-score. Although the difference is not substantial, these results indicate that the RBF kernel is slightly superior in understanding data structures, particularly those that are non-linear and complex, such as the air quality data used in this study.

To strengthen these results, a confusion matrix is used to see the details of the predictions in each class. In the model with the Sigmoid kernel, most of the data was classified correctly (see Figure 5). The Good category was well recognized, although there were still 3 mispredictions in the Moderate I class and 1 misprediction in the Moderate II class. Moderate I, Moderate II, and Unhealthy showed very good accuracy, although Unhealthy had 1 data point that was incorrectly predicted as Good.

Meanwhile, the model with the RBF kernel gives more stable results (see Figure 6). All data in the Moderate I and Moderate II categories were classified perfectly without error, and only one error occurred in the Unhealthy category. The Good category was also classified very well, with only 3 minor errors to the Moderate I class.

3.5 Performance Analysis

Based on the above results, the RBF kernel provides better classification performance. Through the tests conducted, it is evident that the SVM model with the RBF kernel yields superior results compared to the Sigmoid kernel. Although the difference in accuracy is only 0.84%, this difference is sufficient to demonstrate that RBF is better able to recognize complex data patterns, especially in air quality data that exhibit interrelationships between features and tend to be non-linear.

Overall, the model with the RBF kernel also showed a more stable performance during the cross-validation process. This could be because RBF is more flexible in mapping the data projected into a high-dimensional representation, allowing the model to learn better from the variations in the data. In contrast, the Sigmoid kernel appears to be slightly more sensitive to the data distribution and the coef0 parameter value, resulting in more variable prediction results between trials.

These results underscore the importance of selecting the appropriate kernel in constructing classification models. When combined with automated hyperparameter optimization such as Optuna, the combination proved to significantly improve model performance. In the context of air quality classification based on historical sensor data, the RBF kernel is a more reliable and efficient choice than the Sigmoid Kernel. The confusion matrix from modeling analysis can be seen in Figures 4 and 5.

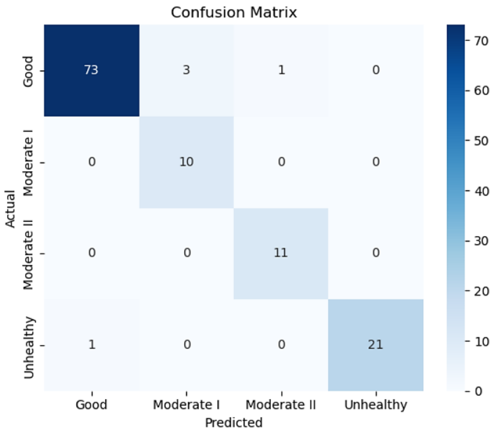

Figure 4. Confusion Matrix of Sigmoid Kernel

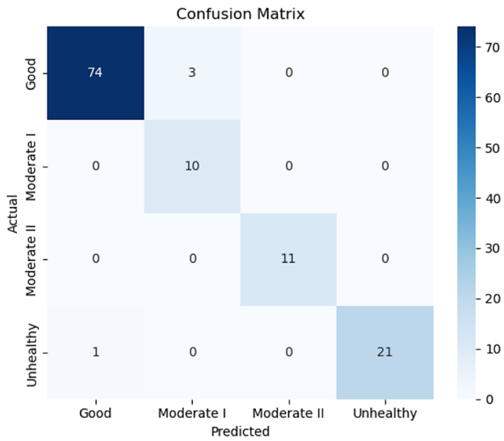

Figure 5. Confusion Matrix of RBF Kernel

3.6 Discussion

The results of this study show that the SVM algorithm, particularly when implemented with the RBF kernel, provides excellent classification performance in determining air quality based on environmental sensor data. The superior results achieved by the RBF kernel are attributed to its ability to manage non-linear and high-dimensional data patterns, which are typical characteristics of environmental parameters such as pollutants, temperature, and humidity. This finding supports earlier studies that reported the RBF kernel's higher stability and accuracy in handling complex environmental and sensor-based data compared to other kernel functions [33], [34].

Model performance was further enhanced by utilizing Optuna as a hyperparameter optimization framework. Unlike conventional tuning methods such as grid search or manual trial-and-error, Optuna applies Bayesian optimization via the TPE algorithm to efficiently identify optimal hyperparameter configurations. This approach allows for more effective parameter exploration and faster convergence toward optimal performance. Previous research also demonstrated that Optuna significantly improves model efficiency and accuracy while reducing computational costs in various machine learning applications [35].

The comparative evaluation between RBF and Sigmoid kernels revealed that the RBF kernel consistently achieved slightly higher accuracy, precision, recall, and F1-score. This suggests that the RBF kernel is better suited for recognizing complex, non-linear interactions among air quality indicators such as CO, HC, TVOC, and eCO_2. However, the Sigmoid kernel still produced strong performance results, indicating its potential suitability for scenarios that prioritize computational simplicity and speed over marginal gains in accuracy.

Overall, the findings underscore the importance of selecting the proper kernel and utilizing automated hyperparameter tuning in developing reliable and efficient air quality classification systems. The combination of SVM and Optuna optimization proved highly effective in enhancing model accuracy and stability. Future research could extend this approach by integrating hybrid ensemble models or deep learning-based feature extraction methods to further improve classification performance and adaptability in real-time air monitoring systems.

4. CONCLUSION

This study successfully achieved its objective of developing an accurate and efficient air quality classification model using the SVM algorithm optimized with Optuna. The results demonstrated that the RBF kernel outperformed the Sigmoid kernel, achieving an accuracy of 96.67% and an F1-score of 96.80%, confirming its superior capability in handling complex and non-linear air quality data. Optuna also played a crucial role in improving model performance by efficiently automating the hyperparameter tuning process, resulting in better accuracy and stability.

Although the findings are promising, this study has certain limitations, particularly the relatively small dataset and the use of a single testing environment, which may limit the model's generalizability. Future research should focus on collecting larger and more diverse datasets, implementing real-time testing in multiple locations, and exploring hybrid or deep learning models to further enhance performance and adaptability. These improvements would support the development of a smarter, based on air quality monitoring system with potential applications in IoT and environmental management.

REFERENCES

[1] D. Hidajat, F. G. Tilana, and I. G. B. S. A. Kusuma, "Dampak Polusi Udara terhadap Kesehatan Kulit," Unram Medical Journal, vol. 12, no. 4, pp. 371–378, Dec. 2023, doi: 10.29303/jk.v12i4.4565.

[2] P. Yin et al., "The effect of air pollution on deaths, disease burden, and life expectancy across China and its provinces, 1990–2017: an analysis for the Global Burden of Disease Study 2017," Lancet Planet Health, vol. 4, no. 9, pp. e386–e398, Sep. 2020, doi: 10.1016/S2542-5196(20)30161-3.

[3] G. Syuhada et al., "Impacts of Air Pollution on Health and Cost of Illness in Jakarta, Indonesia," Int J Environ Res Public Health, vol. 20, no. 4, Feb. 2023, doi: 10.3390/ijerph20042916.

[4] O. Farooq et al., "An enhanced approach for predicting air pollution using quantum support vector machine," Sci Rep, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-69663-2.

[5] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," Neurocomputing, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.

[6] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Leaming, vol. 20, pp. 273–297, 1995, doi: https://doi.org/10.1007/BF00994018.

[7] M. Jabardi, "Support Vector Machines: Theory, Algorithms, and Applications," Infocommunications Journal, vol. 17, no. 1, pp. 66–75, Mar. 2025, doi: 10.36244/ICJ.2025.1.8.

[8] K. L. Du, B. Jiang, J. Lu, J. Hua, and M. N. S. Swamy, "Exploring Kernel Machines and Support Vector Machines: Principles, Techniques, and Future Directions," Mathematics, vol. 12, no. 24, Dec. 2024, doi: 10.3390/math12243935.

[9] M. Noguer i Alonso, "The Mathematics of Support Vector Machines," Oct. 2024.

[10] I. Nur Hermawan and A. Rinaldi Dikananda, "Comparing optimization hyperparameter long short term memory for rainfall prediction model," Institute of Computer Science (IOCS), vol. 16, no. 6, pp. 405–414, Jan. 2025.

[11] Y.-J. Chang, Y.-L. Lin, and P.-F. Pai, "Support Vector Machines with Hyperparameter Optimization Frameworks for Classifying Mobile Phone Prices in Multi-Class," Electronics (Basel), vol. 14, no. 11, p. 2173, May 2025, doi: 10.3390/electronics14112173.

[12] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in Preprint, Jul. 2019. [Online]. Available: http://arxiv.org/abs/1907.10902

[13] M. Conciatori, A. Valletta, and A. Segalini, "Improving the quality evaluation process of machine learning algorithms applied to landslide time series analysis," Comput Geosci, vol. 184, Feb. 2024, doi: 10.1016/j.cageo.2024.105531.

[14] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," Sci Rep, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-56706-x.

[15] A. S. Handayani, S. Soim, T. E. Agusdi, and N. L. Husni, "Air Quality Classification Using Support Vector Machine," Computer Engineering and Applications, vol. 10, no. 1, 2021.

[16] L. Widyarini and H. D. Purnomo, "Air Quality Prediction Using the Support Vector Machine Algorithm," Journal of Information Systems and Informatics, vol. 6, no. 2, pp. 652–661, Jun. 2024, doi: 10.51519/journalisi.v6i2.705.

[17] A. Nandakumar, R. P. Scholar, and A. Professor, "Classification Of Air Pollution Levels Using Supervised Machine Learning Algorithm," Internatinal Journal of Creative Research Thoughts (IJRT), vol. 10, no. 4, pp. 2320–2882, Apr. 2022, [Online]. Available: www.ijcrt.org

[18] N. L. Husni, S. Nurmaini, I. Yani, and A. Silvia, "Intelligent Sensing Using Metal Oxide Semiconductor Based-on Support Vector Machine for Odor Classification," International Journal of Electrical and Computer Engineering (IJECE), vol. 8, pp. 4133–4147, Dec. 2018, doi: 10.11591/ijece.v8i6.pp4133-4147.

[19] M. N. Anshory, M. I. Mazdadi, T. H. Saragih, I. Budiman, and S. W. Saputro, "Application of Adaboost Algorithm with SMOTE and Optuna Techniques in Sleep Disorder Classification," Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 7, no. 2, pp. 415–426, 2025, doi: 10.35882/ijeeemi.v6i3.407.

[20] L. H. Lai et al., "The Use of Machine Learning Models with Optuna in Disease Prediction," Electronics (Switzerland), vol. 13, no. 23, Dec. 2024, doi: 10.3390/electronics13234775.

[21] V. Sinap, "Improving Machine Failure Prediction with Grey Wolf, Whale Optimization, and Optuna Techniques," Gazi University Journal of Science Part A: Engineering and Innovation, vol. 12, no. 1, pp. 154–174, Mar. 2025, doi: 10.54287/gujsa.1544942.

[22] A. Ainul Yaqin, M. Agung Barata, and N. Mahmudah, "Implementation of the Random Forest Algorithm with Optuna Optimization in Lung Cancer Classification," Sistemasi: Jurnal Sistem Informasi, vol. 14, no. 2, pp. 561–569, 2025, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[23] C. M. Rosca, M. Carbureanu, and A. Stancu, "Data-Driven Approaches for Predicting and Forecasting Air Quality in Urban Areas," Applied Sciences (Switzerland), vol. 15, no. 8, Apr. 2025, doi: 10.3390/app15084390.

[24] İ. Aksangür, B. Eren, and C. Erden, "Evaluation of data preprocessing and feature selection process for prediction of hourly PM10 concentration using long short-term memory models," Environmental Pollution, vol. 311, Oct. 2022, doi: 10.1016/j.envpol.2022.119973.

[25] E. Antony, N. S. Sreekanth, R. K. Sunil Kumar, and T. Nishanth, "Data preprocessing techniques for handling time series data for environmental science studies," International Journal of Engineering Trends and Technology, vol. 69, no. 5, pp. 196–207, May 2021, doi: 10.14445/22315381/IJETT-V69I5P227.

[26] D. NURWAHA, "Comparison of kernel functions of support vector machines: A case study for the solar cell output power prediction," International Journal of Energy Applications and Technologies, vol. 7, no. 1, pp. 1–6, Apr. 2020, doi: 10.31593/ijeat.630789.

[27] T. Kee and W. K. O. Ho, “Optimizing Machine Learning Models for Urban Sciences: A Comparative Analysis of Hyperparameter Tuning Methods,” Preprints (Basel), Jul. 2024, doi: 10.20944/preprints202406.0264.v2.

[28] M. Jonas Simanullang, R. Rosnelly, and B. Subhan Riza, “Classification of Big Data Stunting in North Sumatra Using Support Vector Regression Method 1,” Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM), vol. 8, no. 1, pp. 1–10, 2025, doi: 10.24014/ijaidm.v8i1.32177.

[29] A. Tikaningsih, P. Lestari, A. Nurhopipah, I. Tahyudin, E. Winarto, and N. Hassa, “Telematika Optuna Based Hyperparameter Tuning for Improving the Performance Prediction Mortality and Hospital Length of Stay for Stroke Patients,” Telematika, vol. 17, no. 1, pp. 1–16, 2024, doi: 10.35671/telematika.v17i1.2816.

[30] R. Sholehurrohman, I. Valentine, and K. Muludi, “Penilaian Kualitas Udara dan Polusi Menggunakan Algoritma Support Vector Machine,” Jurnal SAINTEKOM, vol. 15, no. 1, pp. 56–68, Mar. 2025, doi: 10.33020/saintekom.v15i1.810.

[31] A. Ainul Yaqin et al., “Sistemasi: Jurnal Sistem Informasi Implementation of the Random Forest Algorithm with Optuna Optimization in Lung Cancer Classification,” Sistemasi, vol. 14, no. 2, pp. 561–569, 2025, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[32] M. Gian, Y. Nita, and A. Solichin, “Early Detection of Hepatitis Disease Using Machine Learning Algorithms,” Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM), vol. 8, no. 3, pp. 545–555, 2025, doi: 10.24014/ijaidm.v8i3.38084.

[33] G. Arisona, A. P. Taruna, D. Irwanto, A. B. Bestari, and W. Juniawan, “Classification Based on the Support Vector Machine for Determining Operational Targets for Controlling Electricity Usage with Conventional Meters: A Case Study of Industrial and Business Tariff Customers from PT PLN (Persero) Indonesia,” IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3529295.

[34] K. B. Dasari and N. Devarakonda, “Detection of TCP-Based DDoS Attacks with SVM Classification with Different Kernel Functions Using Common Uncorrelated Feature Subsets,” International Journal of Safety and Security Engineering, vol. 12, no. 2, pp. 239–249, Apr. 2022, doi: 10.18280/ijsse.120213.

[35] N. Alinda Rahmi and S. Defit, “The Use of Hyperparameter Tuning in Model Classification: A Scientific Work Area Identification,” International Journal on Informatics Visualization, vol. 8, no. 4, pp. 2181–2188, 2024, [Online]. Available: www.joiv.org/index.php/joiv

BIBLIOGRAPHY OF AUTHORS

Nanda Galea was born on July 7, 2003. She is an undergraduate student in the Applied Bachelor’s Program in Electrical Engineering at the State Polytechnic of Sriwijaya (Polsri), class of 2021. With a deep enthusiasm for Internet of Things (IoT), Programmable Logic Controllers (PLC), and machine learning. She is committed to developing real-world skills through hands-on and practical projects. She aspires to become a leading expert in her field by applying innovative technologies to real-life engineering challenges.

Ir. A. Rahman, MT., was born on February 5, 1962. He is a lecturer in the Department of Electrical Engineering at the State Polytechnic of Sriwijaya (Polsri) and has been actively involved in academic teaching and institutional development. He previously served as the Head of the Library Unit at Polsri, where he was responsible for managing library services, collections, and information systems. With a strong educational background in electrical engineering from Sriwijaya University, A. Rahman has made significant contributions to the academic and support infrastructure at Polsri.

Renny Maulidda, S.T., M.T. was born on October 2, 1989. She is a lecturer at the Department of Electrical Engineering, State Polytechnic of Sriwijaya (Polsri), and currently serves as the Head of the Applied Bachelor (D4) Program in Electrical Engineering. She earned her bachelor’s degree from Sriwijaya University and her master’s in Control Engineering and Intelligent Systems from Bandung Institute of Technology (ITB). Her interests include control systems, neural networks, and smart robotics.

Dr. Nyayu Latifah Husni, S.T., M.T. was born on May 3, 1976. She earned her bachelor’s degree in Electrical Engineering from Sriwijaya University. After graduating, she became an assistant lecturer in High Voltage Engineering at the same university. Later, she joined the State Polytechnic of Sriwijaya as a junior lecturer and continued her master’s studies in Electrical Devices at the University of Indonesia. She is actively engaged in teaching and research, with a focus on control engineering, robotics, and artificial intelligence.