Multitek Indonesia: Jurnal Ilmiah
Volume: 20 No. 1 July 2026, Hal. 104 - 115
ISSN: 1907-6223 (print)
ISSN: 2579-3497 (Online)
http://journal.umpo.ac.id/index.php/multitek
doi.org/10.24269/mtkind.v20i1.13841

Analysis Of The Effect Of Gridsearchcv Hyperparameter Tuning Strategy With Variations In Scoring, Test Size, And Random State On Water Quality Classification Performance Using SVM RBF Kernel

Rafli Ardiansyah[1], **Danar Putra Pamungkas**[1], **Made Ayu Dusea Widyadara**[1], **Agoes Santika**[2]
[1]Faculty of Engineering and Computer Science, Informatics Engineering, Universitas Nusantara PGRI Kediri, Kediri City, East Java, Indonesia
[2]Department of Electrical Engincering and Informatics, Universitas Negeri Malang, Malang, East Java, Indonesia
Corresponding Author's Email : rafardiansyahfli@gmail.com

Received : 8 May, 2026 **Revised : 7 June, 2026** **Accepted : 16 June, 2026**

Abstract

The choice of scoring metric in hyperparameter tuning using GridSearchCV is often overlooked, even though this decision directly affects the resulting model's behavior. This study analyzes the impact of scoring metric variations (accuracy, precision, recall), test size (70/30 and 80/20), and random state (0, 21, 42) on the performance of a Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel for water quality classification. The dataset used is a combination of two public datasets from Kaggle totaling 4,259 samples, using three parameters: pH, conductivity, and turbidity, labeled based on World Health Organization (WHO) standards and Indonesian Ministry of Health Regulation No. 32 of 2017. The experiment was designed across 6 scenarios combined with three scoring metrics, resulting in 18 test configurations. Results show that accuracy and precision scoring produce excellent and consistent performance with accuracy ranging from 0.985–0.996 and F1-score of 0.98–1.00 across all scenarios. In contrast, recall scoring causes systematic model degeneration: the model consistently predicts all samples as the positive class, failing to detect any contaminated samples, resulting in accuracy of only approximately 0.50. This phenomenon occurs because GridSearchCV exploits the mathematical definition of recall by selecting parameters that produce a trivial classifier. These findings demonstrate that on a balanced dataset, using recall as the sole scoring metric is counterproductive, and F1-score or accuracy is more recommended as a safe metric for model optimization.

Keywords: *Support Vector Machine, GridSearchCV, Hyperparameter Tuning, Scoring Metric, Water Quality Classification*

INTRODUCTION

When building Machine Learning models, many practitioners treat scoring metric selection in GridSearchCV hyperparameter tuning as a small and unimportant decision. They often just use the default value or pick the same metric as their final evaluation without thinking about the impact[1]. But actually, the scoring metric chosen will directly affect which parameters GridSearchCV picks as the best[2]. In related image-classification studies, Pamungkas et al. reported 79% accuracy for shallot disease classification using SVM [3], while Widyadara et al. combined SVM, GLCM, and Grid Search to achieve 100% accuracy in coffee bean detection[4].

This problem becomes critical when the chosen metric has a one sided mathematical nature, such as recall. Since recall is defined as TP/(TP+FN), its value can be trivially maximized by predicting all samples as the positive class. GridSearchCV, which operates purely based on an objective

function,[5], may find parameter combinations that exploit this definition resulting in a model that appears optimal by metric but lacks genuine classification ability. Although previous studies have shown that SVMs with RBF kernels perform well for water quality classification[6], [7], but none have specifically examined how the choice of evaluation metric affects the tuning process. This study aims to address that gap.

This study systematically analyzes how scoring metric variations accuracy, precision, and recall affect the performance of an SVM RBF kernel model in water quality classification. The analysis was conducted through 18 test configurations combining three scoring metrics, two data splitting ratios (70/30 and 80/20)[8], and three random state values (0, 21, 42) as consistency controls[9]. The dataset used is a combination of two public datasets from Kaggle with three main parameters: pH, conductivity, and turbidity.

RESEARCH METHOD

Before setting up the experiment, previous related studies were reviewed to see where this research stands compared to existing work. The table below shows some studies about water quality classification and SVM with hyperparameter tuning.

Table 1. Comparison of Proposed Model Performance with Previous Studies

Study	Research Objective	Model/Method	Accuracy
GL Pritalia (2022) [6]	Comparative analysis of Mlalgorithms for drinking water quality	6 algorithms (including SVM)	Highest Random Forest (85%)
S. Ramya dkk. (2024) [7]	Comprehensive analysis of multiple classifirers for river water quality	14 algorithms (including SVM)	Highest Gradient Boosting: 99.64%, SVM (99.54%)
M. K. Nallakaruppan dkk. (2024) [10]	Reliable water quality prediction using explainable AI models	RF, DT, Naive Bayes, LR, SVM	Random Forest performed nearly as well (99.9%), while SVM had the lowest performance (63%)
Akash B. Koli dkk. (2025) [11]	Hybrid approach to water quality classification using SVM and XGBoost	SVM, XGBoost	XGBoost (94%) outperforms SVM (64%)
Jauhari Maulani & Mayang Sari (2023) [12]	Comparison of KNN and SVM for water quality classification accuracy	KNN, SVM	SVM (69%) is more accurate than KNN (66%)
Abdul Riqza Ardiansyah dan Danar Putra Pamungkas (2024)[3]	Shallot disease classification using image data	SVM	Highest accuracy: 79%
Muhammad Nur Ichsan and Made Ayu Dusea Widyadara (2025)[4]	Coffee bean detection and non-coffee classification	SVM + GLCM + Grid Search	Accuracy: 100%
Proposed Method	Analyzing the impact of scoring metric selection in GridSreachCV hyperparameter tuning on SVM RBF kernel performance	GridsearchCV-SVM	The optimization achieved an accuracy of 99.6% (80/20 split, R-squared of 21), outperforming the standard SVM

Based on the comparison in Table 1, it can be seen that previous studies focused on algorithm comparison and general model accuracy improvement. Several studies reported that SVM is not always the best algorithm in some cases its performance is far below other algorithms such as Random Forest and XGBoost. However, none of the reviewed studies explicitly analyzed the impact of scoring metric selection in the hyperparameter tuning process on the resulting model behavior. This gap is what differentiates and becomes the main contribution of this research, by proving that GridSearchCV-SVM optimized using the right scoring metric can achieve accuracy up to 99.6%, while also revealing the risk of model degeneration when the chosen scoring metric is asymmetric.

1. Dataset

This study uses two public water quality datasets that were downloaded from Kaggle. Both files were combined into one dataset with 4,259 samples because they share the same parameter structure. Out of all available parameters, this study only uses three main features: pH, conductivity, and turbidity[13], [14] , which gives a balanced class distribution.

Class labeling was performed based on WHO standards[15] and Indonesian Ministry of Health Regulation No. 32 of 2017[16]. Samples were labeled as class 1 (normal) if all three parameters met the standards, and class 0 (contaminated) if any parameter exceeded the limit. The quality standards used are shown in Table 2.

Table 2. Water Quality Parameter Standards

Parameters	Unit	Quality Standard
pH	–	6,5 – 8,5 (WHO & Permenkes RI No. 32/2017)
Conductivity	µS/cm	≤ 750 µS/cm (WHO & Permenkes RI No. 32/2017)
Turbidity	NTU	≤ 5 NTU (WHO)

The resulting class distribution is 51% normal class (2,172 samples) and 49% contaminated class (2,087 samples), so the dataset is considered balanced and does not require any resampling technique.".

2. Eksperiment Design

The experiment was set up with 6 scenarios that combine two data split ratios (70/30 and 80/20) with three different random state values (0, 21, 42)[8],[9]. In each scenario, GridSearchCV was run three times using different scoring metrics accuracy, precision, and recall so the effect of each scoring metric could be compared under the same data conditions [17], [1]. Using three random states makes sure that the results do not rely on just one specific data partition. The experimental design flowchart is shown in Figure 1.

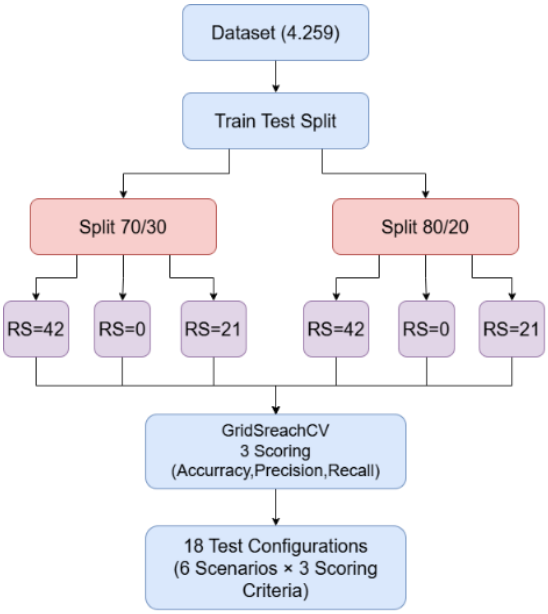

Figure 1. Data Flow Diagram Eksperiment Design

The overall experimental design flow is shown in Figure 1. The dataset of 4,259 samples was first split using Train Test Split with two ratio variations of 70/30 and 80/20. Each ratio was repeated three times using different random states (42, 0, and 21)[8],[9]. producing six unique data partitions [1], All partitions were then processed through GridSearchCV with three different scoring metrics, resulting in 18 test configurations that can be directly compared. All scenario and scoring metric combinations are summarized in Table 3.

Table 3. Experiment Scenarios

Scenario	Split Ratio	Random State	Scoring Metric	Descripstion
1	70/30	42	Accuracy, Precision, Recall	3 configurations
2	70/30	0	Accuracy, Precision, Recall	3 configurations
3	70/30	21	Accuracy, Precision, Recall	3 configurations
4	80/20	42	Accuracy, Precision, Recall	3 configurations
5	80/20	0	Accuracy, Precision, Recall	3 configurations
6	80/20	21	Accuracy, Precision, Recall	3 configurations
total				18 test configurations

From Table 3, every scenario has its own unique combination of split ratio and random state, and each one is tested with the same three scoring metrics accuracy, precision, and recall so we can directly compare how each scoring metric affects the results under the same data conditions [18].

3. Data Preprocessing

Preprocessing was done in four steps in order as shown in Figure 2. The first step is merging and cleaning the data: both datasets were combined, rows with missing values were deleted, and features that were not needed were removed until only three main features were left (pH, conductivity, turbidity)[19]. The second step is assigning class labels based on the rules in Table 2, using WHO standards and Indonesian Ministry of Health Regulation No. 32 of 2017 [15], [16], and split into two labels: normal and contaminated. The third stage is normalization using StandardScaler. Fitting was performed only on the training data, while transformation was applied to both training and testing data [20]. The fourth stage is the final output consisting of data ready to be used in the training process, comprising 4,259 samples with 3 features, 2 class labels, and all values already normalized.

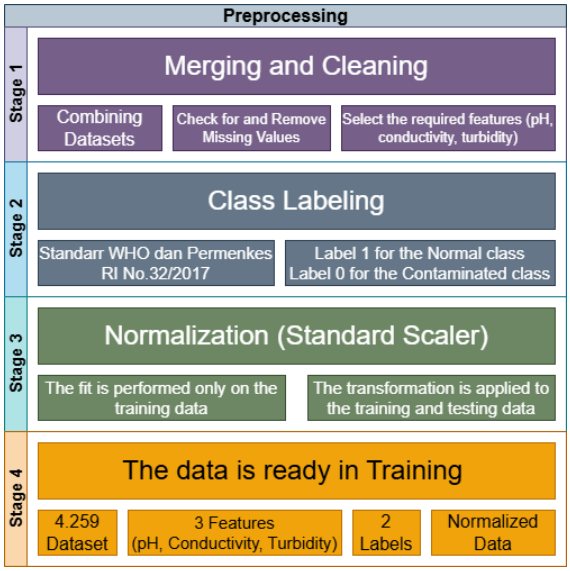

Figure 2. Data Flow Diagram Preprocessing Data

4. Hyperparameter Tuning with GridSearchCV

The search for optimal parameters was conducted using GridSearchCV with 5-fold cross-validation. The parameters tuned were C, gamma, and kernel with the following search space C {0.1; 1; 10; 100; 1000}, gamma {1; 0.1; 0.01; 0.001; 0.0001}, and kernel {rbf, poly, sigmoid}. This combination produces 75 possible configurations tested for each scoring metric in every scenario.[2], [1].

Each GridSearchCV process recorded two kinds of output. The first one is the best parameters (best_params) which covers the values of C, gamma, kernel, and best cross-validation score. The second one is the model performance on test data, which includes accuracy, precision, recall, F1-score, and confusion matrix values (TN, FN, FP, TP) [5].

5. Model Evaluation

The model performance was evaluated using a confusion matrix that gives four basic values: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). From these four values, the following metrics were calculated:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

$$Precision = \frac{(TP)}{(TP + FP)} \quad (2)$$

$$Recall = \frac{(TP)}{(TP + FN)} \quad (3)$$

$$F1 - Score = 2 \times \frac{(Precison \times Recall)}{(Precison + Recall)} \quad (4)$$

In addition to these four metrics, analysis was also performed on the (*best_params*) and (*best cross-validation score*) from each scoring metric to understand the mechanism behind the observed performance differences[18]. Performance comparisons between scenarios are also presented in the form of grouped bar charts to facilitate visual interpretation[21].

RESULTS AND DISCUSSION

1. Dataset

After merging the two datasets from Kaggle, a total of 4,259 valid samples were obtained after removing rows with missing values. Table 4 presents seven representative sample examples along with the value of each parameter and the assigned class label.

Table 4. Dataset Sample Examples

No.	pH	Conductivity (µS/cm)	Turbidity (NTU)	Parameter Excending Standar	Label
1	7,23	432,5	3,1	–	1 (Normal)
2	5,81	1120,3	4,7	Conductivity > 750	0 (Contaminated)
3	8,12	398,0	2,4	–	1 (Normal)
4	6,90	485,2	16,6	Turbidity > 5	0 (Contaminated)
5	9,10	210,0	4,5	pH > 8,5	0 (Contaminated)
6	7,80	676,4	4,2	–	1 (Normal)
7	6,55	1240,0	12,0	Conductivity > 750, Turbidity > 5	0 (Contaminated)

From Table 4, it can be seen that the labeling system is strict and consistent a single parameter exceeding the standard limit is sufficient to classify a sample as contaminated. Sample number 2 is contaminated because conductivity exceeds 750 µS/cm even though pH and turbidity are still within normal limits. Sample number 4 is contaminated because turbidity exceeds 5 NTU, and sample number 5 is contaminated because pH 9.10 exceeds the upper limit of 8.5. Sample number 7 is even contaminated because two parameters simultaneously exceed the standards..

The final class distribution obtained is 51% normal class (2,172 samples) and 49% contaminated class (2,087 samples). This balanced distribution is an important finding because it has direct implications for scoring metric selection on a balanced dataset, accuracy is already a representative and unbiased metric toward either class. This condition also becomes an important context for understanding why using recall as the sole scoring metric is actually counterproductive, as will be discussed in the following subsection.

2. Eksperiment Design

Six experimental scenarios were successfully designed and executed according to the combinations established in Table 3. Each scenario produced three models with different scoring metrics, resulting in a total of 18 test configurations that can be directly compared.

Using three different random state values in each split ratio gave different data partitions that affected model performance. From looking at all scenarios, random state does not always give better performance in a straight line some configurations showed improvement in accuracy, precision, recall, F1-score, and confusion matrix, but others actually showed a decrease. For example, with a 70/30 split ratio, a random state transition from 42 to 0 or 21 produces dynamic results whose direction cannot be fully predicted.

These findings confirm two important things. First, random state does affect model results because it determines the composition of training and test data. Second, although random state triggers variability in the output, it is not the sole determinant that dominates overall model performance the main pattern produced by each scoring metric remains consistent across all scenarios without exception. This proves that the research conclusions do not depend on a single data splitting condition, but are objective and can be generalized [8], [9].

3. Data Preprocessing

In the merging and cleaning step, both datasets from Kaggle were successfully combined and rows with missing values were deleted. The result is a clean dataset with three main features pH, conductivity, and turbidity that is ready for the next step.

In the labeling step, the classification rules were successfully applied consistently to all samples without any ambiguity. This was possible because every parameter has a clear quality limit based on WHO and Indonesian Ministry of Health Regulation No. 32 of 2017 [15], [16]. The final class distribution is 51% normal class (2,172 samples) and 49% contaminated class (2,087 samples), confirming that the dataset is balanced and does not need any resampling technique.

In the normalization step, StandardScaler successfully standardized the scale of all three features which had very different value ranges. Fitting only on training data made sure there was no data leakage, so the evaluation results on test data truly show how well the model generalizes to new data it has never seen before.[22].

After all four preprocessing stages were completed, the final dataset was ready to be used for the training process with a composition of 4,259 samples, 3 normalized features, and 2 balanced class labels.

4. Hyperparameter Tuning With GridSearchCV

Tables 5, 6, and 7 present the optimal parameters best_params and best cross-validation score selected by GridSearchCV for each scoring metric across all 6 scenarios.

Table 5. Optimal Parameters (Scoring: Accuracy)

Sk.	Split	RS	Best Score (CV)	C	Gamma	Kernel
1	70/30	42	0,987	1000	0,1	rbf
2	70/30	0	0,986	1000	1	rbf
3	70/30	21	0,987	1000	1	rbf
4	80/20	42	0,988	1000	0,1	rbf
5	80/20	0	0,988	100	1	rbf
6	80/20	21	0,986	1000	0,1	rbf

From Table 5, GridSearchCV always picked the RBF kernel with high C values (100–1000) in all scenarios. The best cross-validation score was between 0.986–0.988, showing high optimization stability. The fact that RBF kernel was always chosen shows that accuracy gives a clear and stable optimization signal GridSearchCV managed to find a decision boundary that truly separates the two classes.

Table 6. Optimal Parameters (Scoring: Precision)

Sk.	Split	RS	Best Score (CV)	C	Gamma	Kernel
1	70/30	42	0,986	100	1	rbf
2	70/30	0	0,990	1000	1	rbf
3	70/30	21	0,990	1000	0,1	rbf
4	80/20	42	0,990	1000	1	rbf
5	80/20	0	0,990	100	1	rbf
6	80/20	21	0,987	1000	0,1	rbf

The pattern in Table 6 looks very similar to accuracy scoring RBF kernel with high C values (100–1000) was consistently chosen in all scenarios. The best cross-validation score was even a bit higher than accuracy, ranging from 0.986–0.990. The similar pattern between accuracy and precision

scoring shows that on a balanced dataset, both metrics give the same optimization signal and both lead the model to a valid solution..

Table 7. Optimal Parameters (Scoring: Recall)

Sk.	Split	RS	Best Score (CV)	C	Gamma	Kernel
1	70/30	42	1,000	0,1	0,01	poly
2	70/30	0	1,000	0,1	0,01	poly
3	70/30	21	1,000	0,1	0,01	poly
4	80/20	42	1,000	0,1	0,01	poly
5	80/20	0	1,000	0,1	0,01	poly
6	80/20	21	1,000	0,1	0,01	poly

Table 7 shows a completely different and suspicious pattern. Across all 6 scenarios without exception, GridSearchCV always selected the polynomial kernel with C = 0.1 and gamma = 0.01, with a best cross-validation score of 1.000 which appears perfect. This value of 1.000 should serve as a warning signal in practice, a perfect cross-validation score actually indicates exploitation of the objective function, not a truly optimal model. The combination of a very small C = 0.1 and gamma = 0.01 produces a model with excessive regularization that fails to build a meaningful decision boundary, as will be confirmed in the test data evaluation results.

5. Model Evaluation

Table 8 presents the complete evaluation results of the test data for all 18 configurations.

Table 8. Model Performance Evaluation Results on Test Data

Sk.	Split	RS	Scoring	Acc.	Prec.	Rec.	F1	TN/FN/FP/TP
1	70/30	42	Accuracy	0,985	0,99	0,99	0,99	614/11/7/646
			Precision	0,985	0,99	0,99	0,99	614/11/7/646
			Recall	0,514	0,26	0,50	0,34	0/0/621/657
2	70/30	0	Accuracy	0,987	0,99	0,99	0,99	602/11/5/660
			Precision	0,987	0,99	0,99	0,99	602/11/5/660
			Recall	0,525	0,26	0,50	0,34	0/0/607/671
3	70/30	21	Accuracy	0,989	0,99	0,99	0,99	620/11/3/644
			Precision	0,992	0,99	0,99	0,99	620/6/3/649
			Recall	0,512	0,26	0,50	0,34	0/0/623/655
4	80/20	42	Accuracy	0,987	0,99	0,99	0,99	408/6/5/433
			Precision	0,984	0,98	0,98	0,98	408/8/5/431
			Recall	0,515	0,26	0,50	0,34	0/0/413/439
5	80/20	0	Accuracy	0,989	0,99	0,99	0,99	396/4/5/447
			Precision	0,989	0,99	0,99	0,99	396/4/5/447
			Recall	0,529	0,26	0,50	0,35	0/0/401/451
6	80/20	21	Accuracy	0,996	1,00	1,00	1,00	428/3/0/421
			Precision	0,996	1,00	1,00	1,00	428/3/0/421
			Recall	0,497	0,25	0,50	0,33	0/0/428/424

Accuracy and precision scoring gave great and stable performance in all scenarios, with accuracy between 0.985–0.996 and F1-score of 0.98–1.00. Changes in test size and random state did not significantly affect this pattern, confirming that models using both scoring metrics are robust against different data partitions. In contrast, recall scoring produced very poor performance across all scenarios: accuracy of only around 0.50 (equivalent to random guessing), precision of 0.25–0.26, and F1-score of 0.33–0.35. The most critical evidence can be seen in the confusion matrix: the TN value is always 0 across all 6 scenarios. This means the model was unable to detect a single contaminated water sample all samples were predicted as the normal class (class 1).

The consistently observed TN = 0 phenomenon is a direct consequence of GridSearchCV mathematically exploiting the recall objective function. Recall is defined as TP/(TP+FN). The most mathematically efficient way to maximize this value in cross-validation is to predict all samples as the positive class thus FN = 0 and recall = 1.0. GridSearchCV found the parameter combination that produces this behavior: C = 0.1 (excessive regularization), gamma = 0.01 (very small influence range), and polynomial kernel which together produce a hyperplane that fails to separate the two classes.

A comparison of performance across scoring metrics for each scenario is presented in Figures 3–8 as grouped bar charts. Each chart displays four metrics (accuracy, precision, recall, and F1-score) for the three scoring metrics within the same scenario, allowing performance differences to be visualized intuitively.

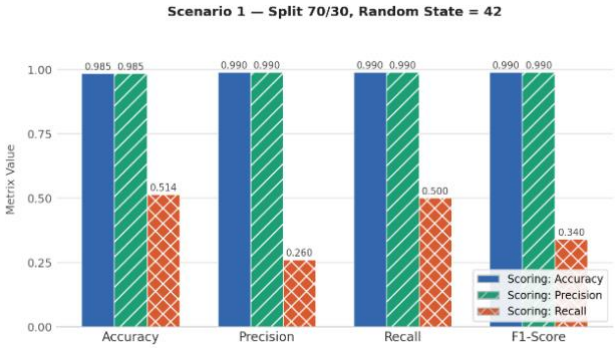

Figure 3. Comparison of Scoring Matrices(scenario 1)

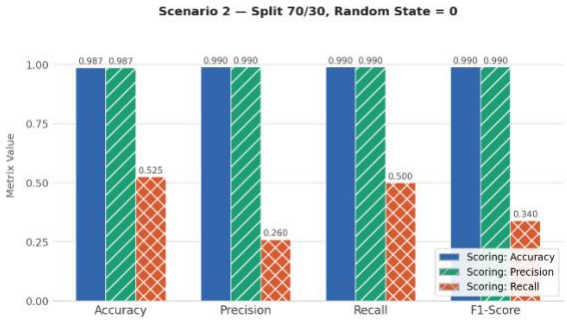

Figure 4. Comparison of Scoring Matrices (skenario 2)

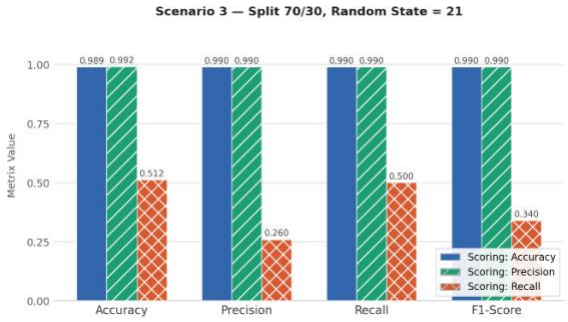

Figure 5. Comparison of Scoring Matrices (skenario 3)

In the group of scenarios with a 70/30 split ratio, a very consistent performance pattern was observed across all three random state values (42, 0, and 21). Accuracy and precision scoring produced excellent performance with accuracy ranging from 0.985–0.989 and F1-score of 0.99 across all scenarios. Both scoring metrics produced identical confusion matrix values in several scenarios, confirming that both direct GridSearchCV toward the same solution. In contrast, recall scoring consistently produced very poor performance with accuracy of only around 0.51–0.52 and F1-score of 0.34 far below the other two scoring metrics. The TN value that is always 0 in all three scenarios proves that the recall-tuned model completely failed to detect contaminated water samples.

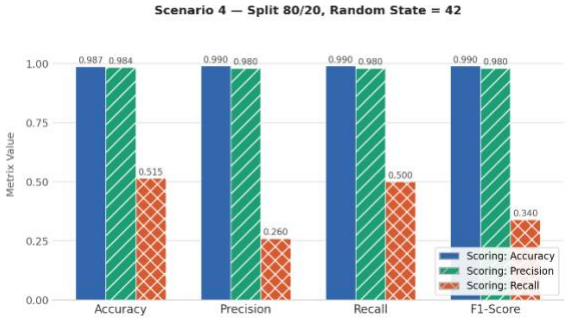

Figure 6. Comparison of Scoring Matrices(scenario 4)

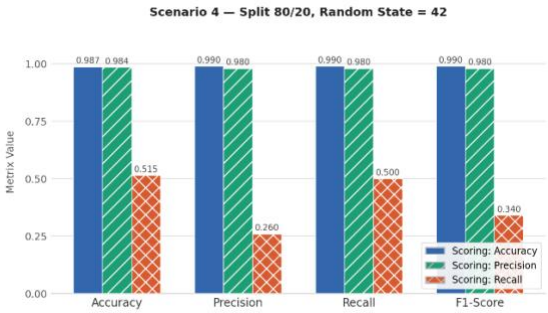

Figure 7. Comparison of Scoring Matrices(scenario 5)

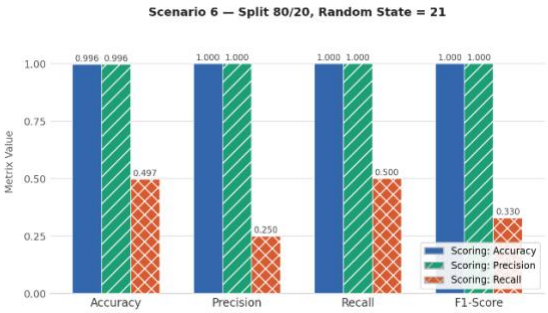

Figure 8. Comparison of Scoring Matrices(scenario 6)

In the group of scenarios with an 80/20 split ratio, the same pattern repeated consistently. Accuracy and precision scoring continued to produce high performance with accuracy of 0.984–0.996 and F1-score of 0.98–1.00. Scenario 6 even achieved a perfect accuracy of 0.996 with both precision and recall at 1.00 the best result among all 18 configurations. Meanwhile, recall scoring again produced model degeneration identical to the 70/30 group TN always 0 and accuracy of only around 0.50.

Overall, changing the split ratio from 70/30 to 80/20 did not change the main pattern produced by each scoring metric. The differences observed were only in small performance value fluctuations due to random state variations, not fundamental changes in model character. This proves that scoring metric is the dominant factor determining model quality, far exceeding the influence of split ratio and random state.

LIMITATION

This study has several limitations. First, only three water quality parameters were used (pH, conductivity, and turbidity), which may not fully represent real-world water quality conditions. Second, the scoring metrics tested were limited to accuracy, precision, and recall. F1-score was not directly tested as a scoring metric in the tuning process. Third, this study only examined SVM with RBF kernel and did not compare it with other machine learning algorithms. Future research is encouraged to address these limitations.

CONCLUSION

This study proves that scoring metric selection in GridSearchCV directly determines the character and reliability of the resulting model. SVM with RBF kernel using accuracy and precision scoring achieved excellent performance across all 6 scenarios, with accuracy of 0.985–0.996 and F1-score of 0.98–1.00. Variations in test size and random state produced dynamic performance fluctuations but did not fundamentally change the main model pattern, confirming that scoring metric is the dominant factor determining model quality.

In contrast, recall scoring caused systematic model degeneration across all scenarios, always producing TN = 0 because GridSearchCV exploited the mathematical definition of recall by selecting

a trivial classifier. This phenomenon occurred regardless of dataset balance, highlighting the risk of using asymmetric metrics as the sole scoring criterion. Based on these findings, accuracy is recommended for balanced datasets while precision serves as an equally reliable alternative. Recall should be avoided as the sole scoring metric. Furthermore, F1-score is theoretically recommended as a promising candidate for future research since it balances both precision and recall, making it mathematically harder for GridSearchCV to exploit trivially.

REFERENCES

[1] E. I. Saputra, M. K. Anam, H. Yenni, H. Hamdani, and A. Zamsuri, "OPTIMALISASI ALGORITMA SUPPORT VECTOR MACHINE PADA ASPECT-. 7, no. 1, pp. 271–279, 2025, Accessed: Jun. 16, 2026. [Online]. Available: https://doi.org/10.31849/zn.v7i1.17800

[2] I. Nurul Hidayah, "OPTIMASI HYPERPARAMETER TUNING MENGGUNAKAN GRIDSEARCHCV PADA METODE RANDOM FOREST DAN SUPPORT VECTOR MACHINE (SVM) UNTUK KLASIFIKASI STATUS INDEKS PEMBANGUNAN MANUSIA DI INDONESIA TAHUN 2022," 2024, Accessed: Jun. 15, 2026. [Online]. Available: http://digilib.unila.ac.id/id/eprint/81889

[3] A. Riqza Ardiansyah and D. Putra Pamungkas, "KLASIFIKASI MENGGUNAKAN METODE SUPPORT VECTOR MACHINE UNTUK MENDETEKSI PENYAKIT TANAMAN BAWANG MERAH," *Jurnal Nusantara Of Engineering*, vol. 7, 2024, [Online]. Available: https://ojs.unpkediri.ac.id/index.php/noe

[4] M. Biji Kopi, M. Nur Ichsan, M. Ayu Dusea Widyadara, and U. Mahdiyah, "Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi) 2025 1024 P emanfaatan Support Vector Machine dalam," Online, 2025. Accessed: Jun. 15, 2026. [Online]. Available: https://doi.org/10.29407/75j7ar20

[5] I. W. A. Sparta and P. E. Parianthana, "Support Vector Machine Performance In Diabetes Data Classification With GridsearchCV Optimization," *Jurnal Multidisiplin Sahombu*, vol. 5, no. 01, pp. 179–186, 2025, Accessed: Jun. 16, 2026. [Online]. Available: https://ejournal.seaninstitute.or.id/index.php/JMS/article/view/6215

[6] G. L. Pritalia, "Analisis Komparatif Algoritme Machine Learning pada Klasifikasi Kualitas Air Layak Minum," 2022. Accessed: Jun. 15, 2026. [Online]. Available: doi: 10.24002/konstelasi.v2i1.5630

[7] S. Ramya, S. Srinath, and P. Tuppad, "Comprehensive analysis of multiple classifiers for enhanced river water quality monitoring with explainable AI," *Case Studies in Chemical and Environmental Engineering*, vol. 10, Dec. 2024, doi: 10.1016/j.cscee.2024.100822.

[8] A. Rácz, D. Bajusz, and K. Héberger, "Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification," *Molecules*, vol. 26, no. 4, Feb. 2021, doi: 10.3390/molecules26041111.

[9] T. Horváth, R. G. Mantovani, and A. C. de Carvalho, "Effects of random sampling on svm hyper-parameter tuning," in *International Conference on Intelligent Systems Design and Applications*, Springer, 2016, pp. 268–278. Accessed: Jun. 16, 2026. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-53480-0_27

[10] M. K. Nallakaruppan, E. Gangadevi, M. L. Shri, B. Balusamy, S. Bhattacharya, and S. Selvarajan, "Reliable water quality prediction and parametric analysis using explainable AI models," *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-56775-y.

[11] A. B. Koli, B. Faijan, S. Akil, B. Kantilal, D. Madhukar, and P. R. Sanjay, "A Hybrid Approach to Water Quality Classification Using SVM and Xgboost Method", doi: 10.51244/IJRSI.

[12] J. Maulani and M. Sari, "Komparasi Metode K-Nearest Neighbor (Knn) Dengan Support Vector Machine (Svm) Terhadap Tingkat Akurasi Klasifikasi Kualitas Air," 2023. Accessed: Jun. 16, 2026. [Online]. Available: DOI:10.30591/smartcomp.v12i2.4205

[13] Stacyana Jesika, Suci Ramadhani, and Yohanna Permata Putri, "Implementasi Model Machine Learning dalam Mengklasifikasi Kualitas Air," *Jurnal Ilmiah Dan Karya Mahasiswa*, vol. 1, no. 6, pp. 382–396, Nov. 2023, doi: 10.54066/jikma.v1i6.1162.

[14] M. Junus and F. Abdillah, *PENERAPAN METODE SUPPORT VECTOR MACHINE UNTUK KLASIFIKASI KUALITAS AIR MINUM BERBASIS IOT*. 2025. Accessed: Jun. 16, 2026. [Online]. Available: https://www.researchgate.net/publication/389945680_PENERAPAN_METODE_SUPPORT_VECTOR_MACHINE_UNTUK_KLASIFIKASI_KUALITAS_AIR_MINUM_BERBASIS_IOT

[15] World Health Organization, *Guidelines for Drinking-water Quality*, 4th edition. Geneva: WHO press, 2011.

[16] Kementerian Kesetahan Republik Indonesia, "Peraturan Menteri Kesehatan Republik Indonesia Nomor 32 Tahun 2017 tentang Standar Baku Mutu Kesehatan Lingkungan," Jakarta, 2017.

[17] N. P. E. M. Anggarini and A. Muliantara, "Memprediksi Kelulusan Mahasiswa Graduate dan Dropout dengan Support Vector Machine dan GridSearchCV," *Jurnal Nasional Teknologi Informasi dan Aplikasinya*, vol. 2, no. 3, pp. 475–480, 2024, Accessed: Jun. 16, 2026. [Online]. Available: DOI: https://doi.org/10.24843/JNATIA.2024.v02.i03.p04

[18] F. Dewi, N. C. H. Wibowo, M. R. Handayani, and K. Umam, "Evaluasi Hyperparamter Tuning Pada Support Vector Machine (Svm) Dalam Klasifikasi Ulasan Hotel Di Tripadvisor," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 3, pp. 2584–2593, 2025, Accessed: Jun. 16, 2026. [Online]. Available: DOI: https://doi.org/10.29100/jipi.v10i3.7774

[19] N. Nuraeni, "Klasifikasi dengan metode Support Vector Machine (SVM) dalam menentukan kualitas air minum," *Maliki Interdisciplinary Journal*, vol. 3, no. 6, pp. 2248–2258, 2025, Accessed: Jun. 16, 2026. [Online]. Available: https://urj.uin-malang.ac.id/index.php/mij/article/view/14528

[20] N. W. A. Prasetya, L. P. Wanti, R. Purwanto, I. Bahroni, and R. Listyaningrum, "Evaluasi Kinerja Model Machine Learning dalam Klasifikasi Penyakit THT: Studi Komparatif NaÃ¯ ve Bayes, SVM, dan Random Forest," *Infotekmesin*, vol. 16, no. 2, pp. 312–320, 2025, Accessed: Jun. 16, 2026. [Online]. Available: DOI: https://doi.org/10.35970/infotekmesin.v16i2.2798

[21] "Evaluation Metrics in Machine Learning - GeeksforGeeks." Accessed: Nov. 04, 2025. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/metrics-for-machine-learning-model/

[22] D. Tao and X. Tang, "Random sampling based SVM for relevance feedback image retrieval," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, IEEE, 2004, pp. II–II. Accessed: Jun. 16, 2026. [Online]. Available: DOI: 10.1109/CVPR.2004.1315225