Jurnal Multidisiplin Sahombu
Volume 5, Number 01, 2025, DOI 10.58471/jms.v5i01
ESSN 2809-8587 (Online)
https://ejournal.seaninstitute.or.id/index.php/JMS/index

Support Vector Machine Performance In Diabetes Data Classification With GridsearchCV Optimization

I Wayan Adi Sparta[1*], Putu Eka Parianthana[2]
[1*,2]Prodi Sistem Informasi, Fakultas Teknologi dan Ilmu Kesehatan, Universitas Bali Dwipa

Article Info
Keywords:
Support Vector Machine, Diabetes, Classification, Machine Learning, Parameter Optimization

ABSTRACT
Diabetes is one of the most common chronic diseases and has the potential to cause serious complications if not detected early. This study evaluates the performance of the Support Vector Machine (SVM) algorithm in diabetes data classification using parameter optimization with GridSearchCV. The dataset used includes eight health features from Kaggle, which include factors such as glucose levels, blood pressure, and body mass index (BMI). After parameter optimization, the SVM model achieved 74% accuracy on the test data, with precision and recall performance varying between classes. This research highlights the importance of hyperparameter optimization in improving the accuracy and balance of health data classification. These results provide insight into the application of machine learning technologies to support early detection and medical decision-making.

Corresponding Author:
I Wayan Adi Sparta
Prodi Sistem Informasi, Fakultas Teknologi dan Ilmu Kesehatan, Universitas Bali Dwipa
adisparta11@gmail.com

INTRODUCTION

Diabetes is a global health threat with prevalence increasing every year. Early detection of diabetes is essential to prevent serious complications and reduce the public health burden. In an effort to support early diagnosis, machine learning technologies such as Support Vector Machine (SVM) have been widely used to analyze health data (Afifuddin & Hakim, 2023; Khan et al., 2022; Lin, 2024; Purbolaksono et al., 2021) .

SVM is a model-based learning algorithm that aims to optimally separate data from different classes with a specific hyperplane (Aini et al., 2024; Khadangi & Bagheri, 2013; Valero-Carreras et al., 2023) . However, SVM performance is strongly influenced by parameters such as regulation (C), kernel parameters (gamma), and the type of kernel used. For this reason, hyperparameter optimization is an important step to improve model performance. One widely used method is GridSearchCV (Hidayatulloh et al., 2024; Ishlah et al., 2023) , which allows exploration of various parameter combinations to determine the best configuration based on the evaluation of certain metrics.

This study aims to evaluate the performance of SVM optimized with GridSearchCV in diabetes data classification. The focus is on accuracy, precision, recall, and F1-score to provide a comprehensive analysis of the model's capabilities.

Jurnal Multidisiplin Sahombu
Volume 5, Number 01, 2025, DOI 10.58471/jms.v5i01
ESSN 2809-8587 (Online)
https://ejournal.seaninstitute.or.id/index.php/JMS/index

METHODS

Datase

The dataset used in this analysis is taken from Kaggle, which includes eight key health features and one binary label as output. These features include number of pregnancies, plasma glucose level, diastolic blood pressure, skinfold thickness, serum insulin level, body mass index (BMI), family history of diabetes (DiabetesPedigreeFunction), and age. The binary label on the dataset indicates the result of the diagnosis, which is 0 for no diabetes and 1 for diabetes.

Preprocessing

At this stage, the data is examined to find missing values and anomalies. Next, the features are normalized to ensure a more uniform distribution of the data, which is important for improving the performance of the model (Suryawan et al., 2024) . After preprocessing is completed, the dataset is divided into two parts, 80% for training data and 20% for testing data.

Hyperparameter Optimization

This analysis is used to find the best combination of model parameters, including the C value that represents regulation, the gamma value that determines the influence of a data point, and the type of kernel used (in this case Radial Basis Function or RBF). The parameter combinations tested include C of 1, 10, and 100, gamma of 0.1, 0.01, and 0.001, and rbf kernel.

Evaluation

After optimization is complete, the SVM model is trained using the best parameter combination obtained. Model performance is evaluated based on accuracy, precision, recall, and F1-score metrics. In addition, confusion matrix is used to provide a more in-depth analysis of the model's ability to classify the data.

Analysis Flow

This process starts from preparing disease data, namely diabetes, then preprocessing data such as seeing if there is missing data, normalizing data or standardizing features. Then the data is divided into two parts, namely training data and testing data. Next, initialization of Support Vector Machine models such as RBF or linear and determination of initial parameters for Support Vector Machine such as C and gamma. The next step is Hyperparameter Optimization with GridSearchCV to perform a grid search on various combinations of hyperparameters for SVM, such as C, gamma, and kernel. Then the Support Vector Machine model is trained using the training data and the best hyperparameters from GridSearchCV. Model evaluation is also carried out to make predictions using the trained model to predict diabetes labels on test data and Performance Evaluation to measure model performance using evaluation metrics such as accuracy, precision, recall and F1-score. The final stage of this analysis is to draw conclusions from the analysis that has been carried out.

Jurnal Multidisiplin Sahombu
Volume 5, Number 01, 2025, DOI 10.58471/jms.v5i01
ESSN 2809-8587 (Online)
https://ejournal.seaninstitute.or.id/index.php/JMS/index

RESULTS AND DISCUSSION

Dataset

Data taken from the Kaggle.com website, with a total of 768 data, with parameters Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, Diabates Predigree Function and Age.

Table 1. Diabetes Disease Dataset

No.	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes PedigreeFunction	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0	0.232	54	1

Support Vector Machine Classification Results with Optimization

```
Fitting 5 folds for each of 32 candidates, totalling 160 fits
({'C': 100, 'gamma': 0.001, 'kernel': 'rbf'},
 0.7445887445887446,
 array([[132,  18],
        [ 41,  40]]),
 '              precision    recall  f1-score   support\n\n           0       0.76
0.88      0.82       150\n           1       0.69      0.49      0.58        81\n\n
accuracy                           0.74       231\n   macro avg       0.73      0.69
0.70       231\nweighted avg       0.74      0.74      0.73       231\n')
```

Figure 2. SVM output

Table 2. SVM Model Classification Results

Class	Precision	Recall	F1-Score	Support
No Diabetes (0)	0.76	0.88	0.82	150
Diabetes (1)	0.69	0.49	0.58	81
Overall	0.74	0.74	0.73	231

Best Parameters

1. C: 100
   This parameter function controls the tolerance to error in the SVM model. A larger value such as 100 means the model is more focused on classifying all the data correctly.
2. gamma: 0.001
   The Gamma function determines how far away the influence of a data point is. The resulting value of 0.001 makes the model consider the influence of more distant data points to form a decision, which is suitable for this dataset.

Jurnal Multidisiplin Sahombu
Volume 5, Number 01, 2025, DOI 10.58471/jms.v5i01
ESSN 2809-8587 (Online)
https://ejournal.seaninstitute.or.id/index.php/JMS/index

3. kernel: rbf (Radial Basis Function)
   This Kernel code function works by creating a more complex feature space to separate data that cannot be linearly separated.

Model Performance

Accuracy on Test Data is 74.46%, with the Model being able to correctly classify about 74 out of 100 data on the test set.

Confusion Matrix

First row (class 0 - No Diabetes):

1. 132 (True Negatives): The model successfully predicted 132 people who did not have diabetes.
2. 18 (False Positives): The model incorrectly predicted 18 people as diabetic, when in fact they did not have diabetes.

Second row (class 1 - Diabetes):

1. 40 (True Positives): The model successfully predicted 40 people who did have diabetes.
2. 41 (False Negatives): The model incorrectly predicted 41 people as not having diabetes, when they actually had diabetes.

In Class 0 (Not Diabetic): Precision (76%) of all classifications that called someone not diabetic, 76% were correct. Recall (88%) of all people who were actually not diabetic, 88% were successfully detected. F1-Score (82%) The combination of precision and recall shows a fairly good performance. While in Class 1 (Diabetes): Precision (69%) of all predictions that called someone diabetic, 69% were correct. Recall (49%) of all people who were actually diabetic, only 49% were successfully detected. F1-Score (58%): This performance shows that the model is not yet optimal in detecting diabetes cases.

The overall result is Accuracy: 74%, with Macro Avg (Average per Class): Precision: 73%, Recall: 69%, F1-Score: 70%. And Weighted Avg: Precision: 74%, Recall: 74%, F1-Score: 73%. This analysis shows that the use of the Support Vector Machine model with GridSearchCV optimization in diabetes classification is able to achieve maximum accuracy results of 0.74 or 74%. This result shows that the Support Vector Machine model has greater potential in classifying diabetes or relevant health data. The accuracy of 74% reflects the success rate of the model in classifying diabetic patients correctly.

Jurnal Multidisiplin Sahombu
Volume 5, Number 01, 2025, DOI 10.58471/jms.v5i01
ESSN 2809-8587 (Online)
https://ejournal.seaninstitute.or.id/index.php/JMS/index

SVM Data Visualization

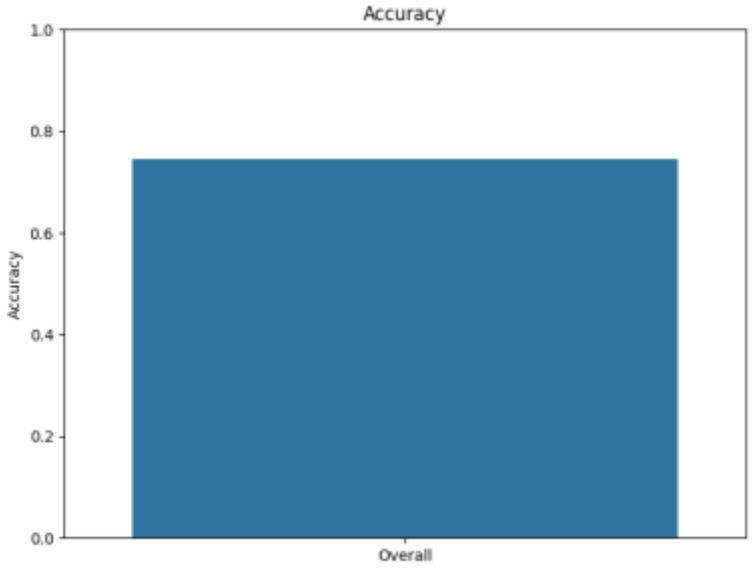


Figure 3. SVM Accuracy Result Diagram

The figure above shows the overall accuracy value of the classification model used because accuracy reflects the overall performance of the model not per class, with results reaching around 74%. This means that the model successfully made correct predictions on 74% of the total test data.

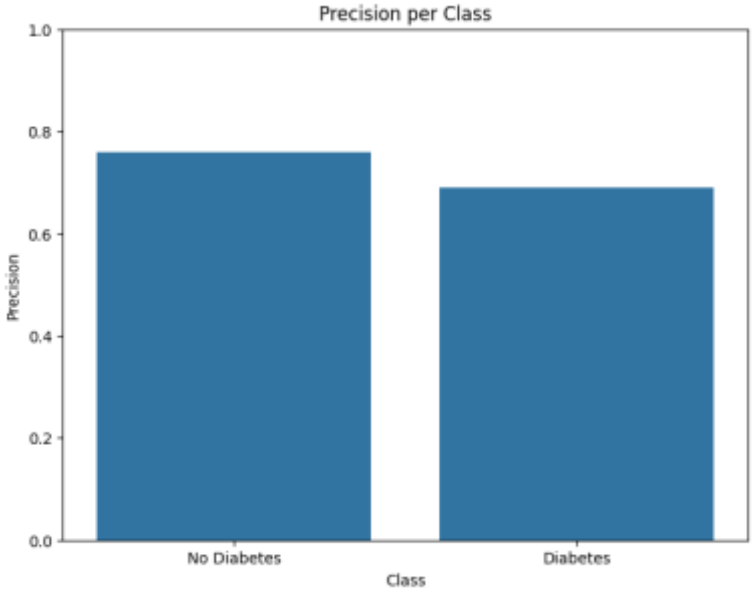


Figure 4. SVM Precision Result Diagram

In the figure above, we can see the results of SVM Precission. Precision is a measure that shows how many positive examples are correctly predicted from the total predicted positive examples. It can be seen that No Diabetes with Precission 0.76 or 76% means that the prediction of "No Diabetes" really does not have diabetes higher than Diabetes states that

Jurnal Multidisiplin Sahombu
Volume 5, Number 01, 2025, DOI 10.58471/jms.v5i01
ESSN 2809-8587 (Online)
https://ejournal.seaninstitute.or.id/index.php/JMS/index

the model is good enough in classifying people who do not have diabetes compared to people who have diabetes.

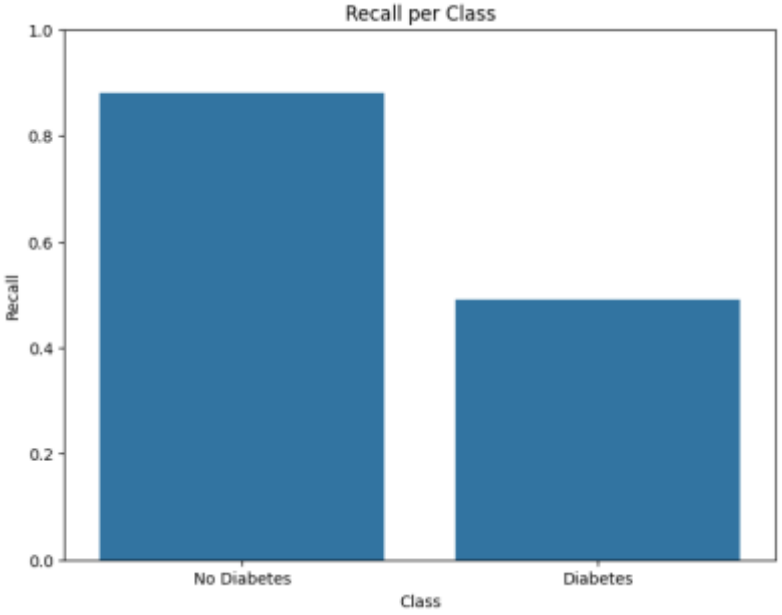
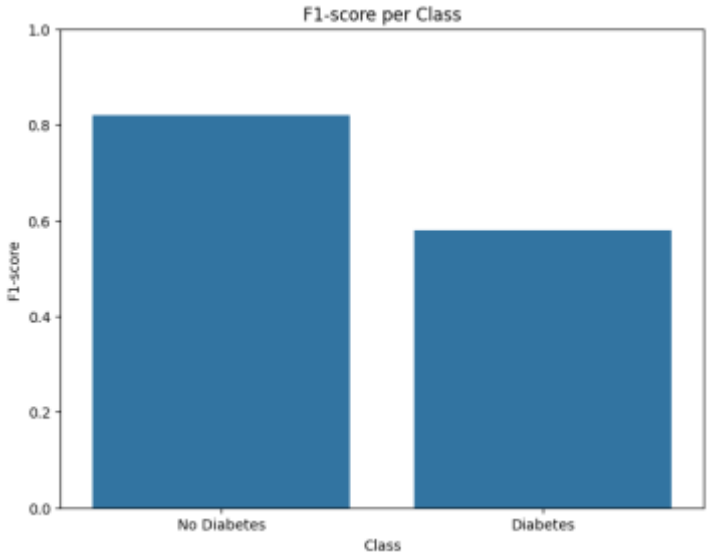


Figure 5. SVM Recall Result Diagram

In the figure above can be seen the results of the SVM Recall results. Recall is a measure that shows how many positive examples are correctly predicted from the total positive examples. Recall shows how good the model is in cases that actually exist (for example, finding people who are really diabetic) The Recall result is 0.88 or 88% of people who are really not diabetic can be found correctly by the model then Recall 0. 49 or 49% of people who are really diabetic can be found by the model. With this, the performance of SVM in identifying non-diabetic people, but less good in finding people who are actually diabetic.



F1-Score SVM Result Diagram

Jurnal Multidisiplin Sahombu
Volume 5, Number 01, 2025, DOI 10.58471/jms.v5i01
ESSN 2809-8587 (Online)
https://ejournal.seaninstitute.or.id/index.php/JMS/index

In the figure above, we can see the results of the SVM F1-Score results. F1-Score is a measure that shows the balance between precision and recall. No Diabetes F1-Score 0.82 or 82% shows the model works quite balanced between precision and balance in this class. While Diabetes F1-Score is 0.58 or 58%, this shows that although the recall for diabetes is low, the precision is quite good. From these results, the model performance is more balanced in recognizing people who are not diabetic (No Diabetes), but there is a big deficiency in recognizing people who are truly diabetic.

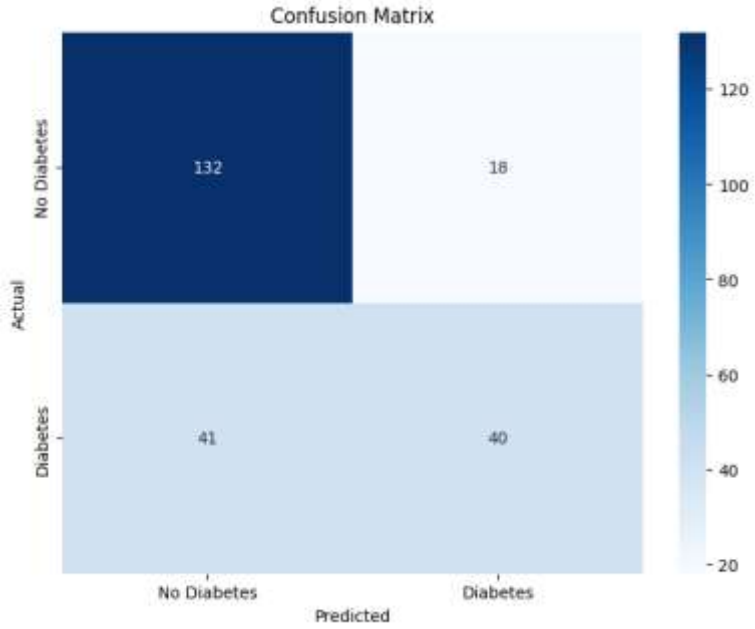


Figure 7. SVM Confusion Matrix Result Diagram

In the figure above is the result of the SVM Confusion Matrix which is used to evaluate the performance of the classification model. In the first row (No Diabetes) there are 132 The amount of actual No Diabetes data that was correctly classified by the model (True Negative, TN) then 18: The amount of actual No Diabetes data that was incorrectly classified as Diabetes by the model (False Positive, FP). In the second row (Diabetes) there are 41 the number of true data of Diabetes that are wrongly classified as No Diabetes by the model (False Negative, FN) then 40 the number of true data of Diabetes that are correctly classified by the model (True Positive, TP).

CONCLUSION

Based on the results of the analysis carried out, it can be concluded that the performance of Support Vector Machine in Diabetes Data Classification with GridSearchCV Optimization provides good results. The use of Support Vector Machine (SVM) optimized with GridSearchCV succeeded in providing 74% accuracy in the classification of diabetes data. These results show that the SVM method has great potential to be applied in health data analysis, including early detection of diabetes. However, the model performance can still be improved, especially in the recall aspect for detecting positive cases of diabetes (class 1),

which only reached 49%. Overall, the combination of precision, recall, and F1-score showed adequate performance for the category "No Diabetes," but still requires improvement for the category "Diabetes." This study confirms the importance of parameter optimization and further testing to improve classification performance.

REFERENCE

Afifuddin, A., & Hakim, L. (2023). Deteksi Penyakit Diabetes Mellitus Menggunakan Algoritma Decision Tree Model Arsitektur C4. 5. *Jurnal Krisnadana*, *3*(1), 25–33.

Aini, W. R., Sudipa, I. G. I., Sandana, I. P. D., Putra, D. M. D. U., & Indrawan, I. G. A. (2024). IDENTIFYING FAKE ACCOUNTS IN SOCIAL MEDIA COMMERCIAL VIDEOS USING SUPPORT VECTOR MACHINE METHOD. *Proceeding International Conference on Information Technology, Multimedia, Architecture, Design, and E-Business*, *3*, 79–86.

Hidayatulloh, F., Andika, I. G., Suryawan, I. W. D., Ariantini, M. S., & Sudipa, I. G. I. (2024). Analisis Sentimen Opini Publik Terhadap Konser Coldplay Di Jakarta Pada Twitter Menggunakan Metode Support Vector Machine. *INFORMAL: Informatics Journal*, *9*(2), 137–144.

Ishlah, A. W., Sudarno, S., & Kartikasari, P. (2023). IMPLEMENTASI GRIDSEARCHCV PADA SUPPORT VECTOR REGRESSION (SVR) UNTUK PERAMALAN HARGA SAHAM. *Jurnal Gaussian*, *12*(2), 276–286. https://doi.org/10.14710/j.gauss.12.2.276-286

Khadangi, E., & Bagheri, A. (2013). Comparing MLP, SVM and KNN for predicting trust between users in Facebook. *Proceedings of the 3rd International Conference on Computer and Knowledge Engineering, ICCKE 2013*, *Iccke*, 466–470. https://doi.org/10.1109/ICCKE.2013.6682864

Khan, A., Khan, A., Khan, M. M., Farid, K., Alam, M. M., & Su'ud, M. B. M. (2022). Cardiovascular and Diabetes Diseases Classification Using Ensemble Stacking Classifiers With SVM as a Meta Classifier. *Diagnostics*, *12*(11), 2595. https://doi.org/10.3390/diagnostics12112595

Lin, A. K. (2024). The AI Revolution in Financial Services: Emerging Methods for Fraud Detection and Prevention. *Jurnal Galaksi*, *1*(1), 43–51. https://doi.org/10.70103/galaksi.v1i1.5

Purbolaksono, M. D., Irvan Tantowi, M., Imam Hidayat, A., & Adiwijaya, A. (2021). Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(2), 393–399. https://doi.org/10.29207/resti.v5i2.3008

Suryawan, I. G. T., Pratiwi, N. L. S., Sudipa, I. G. I., & Anandita, I. B. G. (2024). Performance of Moving Average and Exponential Smoothing Methods in Forecasting Demand for Blood Components. *ISAR Journal of Science and Technology*, *2*(11), 7–17. https://doi.org/10.5281/zenodo.14250561

Valero-Carreras, D., Alcaraz, J., & Landete, M. (2023). Comparing two SVM models through different metrics based on the confusion matrix. *Computers & Operations Research*, *152*, 106131.