Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2035−2045
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8880

Classification of Diabetes Diseases Based on Medical Features Using Optimized Support Vector Machine

Ita Arfyanti[1,*], Amelia Yusnita[2], Pitrasacha Adytia[1]

[1] Program Studi Sistem Informasi, STMIK Widya Cipta Dharma, Samarinda, Indonesia
[2] Program Studi Teknik Informatika, STMIK Widya Cipta Dharma, Samarinda, Indonesia
Email: [1,*]ita@wicida.ac.id, [2]amelia@wicida.ac.id, [3]pitra@wicida.ac.id
Corresponding Author Email: ita@wicida.ac.id
Submitted: **08/12/2025**; Accepted: **31/12/2025**; Published**: 31/12/2025**

Abstract–Diabetes mellitus is a chronic disease caused by impaired glucose metabolism and has become a global health threat with a steadily increasing prevalence each year. According to WHO and IDF, the number of people living with diabetes is projected to reach 783 million by 2045. This condition demands the development of an accurate and efficient early detection system to support medical decision-making. This study aims to develop an optimized Support Vector Machine (SVM)-based classification model to enhance the accuracy and interpretability of diabetes prediction. The dataset used is the Pima Indians Diabetes Dataset, which consists of eight medical features such as glucose level, blood pressure, and body mass index (BMI). The research stages include data preprocessing, class balancing using the Synthetic Minority Over-sampling Technique (SMOTE), parameter optimization with GridSearchCV, and interpretability analysis through SHapley Additive exPlanations (SHAP). The results show that the optimized SVM model with the Radial Basis Function (RBF) kernel achieved an accuracy of 82%, with a significant improvement in the diabetes class recall value from 0.564 to 0.83 after optimization. The Area Under Curve (AUC) value of 0.871 indicates the model's effectiveness in distinguishing between positive and negative classes. The SHAP analysis reveals that Glucose, Age, BMI, and Diabetes Pedigree Function are the most influential features in prediction. These findings emphasize that the combination of normalization, balancing, hyperparameter optimization, and interpretability produces a reliable and transparent SVM model. This model has strong potential for implementation in Clinical Decision Support Systems (CDSS) for accurate and explainable early diabetes detection.

Keywords: Support Vector Machine (SVM); Hyperparameter Optimization; SMOTE; SHAP; Diabetes Classification

1. INTRODUCTION

Diabetes mellitus is one of the most significant chronic diseases globally, caused by disorders of glucose metabolism, where blood sugar levels increase due to impaired insulin production or function[1][2][3]. The World Health Organization (WHO) estimates that the number of people with diabetes will rise to more than 643 million by 2030, making it a serious global health threat[4][5]. This condition leads to various complications such as heart disease, kidney failure, blindness, and limb amputations if not detected early[6]. In this context, the development of automated classification systems based on machine learning becomes crucial to support early diagnosis and medical decision-making[7]. With the availability of large and diverse medical data, Data Mining and Machine Learning methods such as Support Vector Machine (SVM) have emerged as potential approaches for accurately classifying diabetes[8].

The high global prevalence of diabetes poses a major challenge to modern healthcare systems. According to the International Diabetes Federation (IDF), there were over 537 million people living with diabetes in 2021, and this number is projected to increase to 783 million by 2045[9]. In Indonesia alone, the prevalence of diabetes reached 10.6% of the total adult population, positioning it as one of the countries with the fastest-growing number of diabetes cases in Southeast Asia[10][11]. The primary issues faced are delayed diagnosis due to difficulties in interpreting complex and varied medical data. Furthermore, class imbalance in medical datasets also leads to decreased accuracy of predictive models [8]. Therefore, an approach capable of optimizing classification by considering data complexity and interrelated medical variables is required.

To address these issues, the optimized Support Vector Machine (SVM) method presents an effective solution due to its capability in handling high-dimensional and non-linear data[12]. SVM operates by finding the optimal hyperplane that maximally separates data between classes, making it highly suitable for classifying diseases with complex characteristics like diabetes[13][14]. Optimization of SVM can be performed using techniques such as the Firefly Algorithm, Particle Swarm Optimization, or Backward Elimination to enhance accuracy and computational efficiency[6]. The application of these optimization techniques has been proven to yield classification models with accuracy rates above 95%, significantly improving early diabetes detection capabilities[9].

Several previous studies have demonstrated the effectiveness of the SVM method in diabetes disease classification. Research by [12] utilizing a combination of SVM with Backward Elimination achieved an accuracy of 85.71% on the PIMA dataset. Meanwhile, [15] implemented SVM+PSO and SVM with a POLY kernel, obtaining a best result of 83.99% using PSO. Another study by [6] utilized the Firefly optimization algorithm to enhance SVM performance, achieving an accuracy of 98.45%. Additionally, [9] found that the Radial Basis Function (RBF) kernel yielded the best results with an accuracy of up to 97% in detecting diabetes. These four studies reinforce the position of SVM as a leading algorithm in data-based medical disease classification analysis.

Nonetheless, two main challenges persist: class imbalance in diabetes datasets often reduces model sensitivity towards positive patients, and the lack of prediction interpretability makes the model difficult to rely upon as a clinical decision support system. To overcome these problems, this research proposes the use of hyperparameter-optimized SVM

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2035−2045
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8880

(such as Grid Search), combined with balancing techniques like SMOTE, and explainable AI analysis such as SHAP, so that the model is not only accurate but also explainable.

The objectives of this research are to develop and evaluate an optimized SVM model for diabetes classification that can improve performance metrics (e.g., accuracy, sensitivity), while also mapping the contribution of medical features (such as glucose level, BMI, blood pressure, age) to the prediction through SHAP interpretation. It is hoped that the results will provide a scientific basis for the application of a clinical decision support system (CDSS) in the early detection of diabetes, with potential for real-world implementation in healthcare facilities.

2. RESEARCH METHODOLOGY

This research was conducted to develop and evaluate a Support Vector Machine (SVM) model for classifying diabetic patients based on medical parameters. Each stage was designed to enable the model to achieve high accuracy and to be clinically explainable through feature interpretability analysis. Prior to detailing each step, Figure 1 illustrates the overall research methodology workflow. This workflow consists of five main stages, namely: (1) Data Acquisition and Pre-processing, (2) Feature Analysis, (3) SVM Parameter Optimization, (4) Model Training and Validation, and (5) Evaluation and Interpretation of Results.

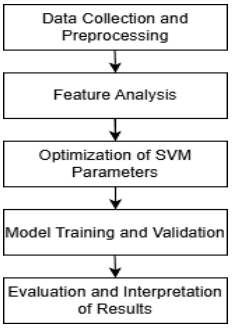


Figure 1. Research Stages

This study aims to develop and evaluate a Support Vector Machine (SVM) model for diabetes classification based on medical parameters, through a structured workflow based on Figure 1 above. The process begins with Data Acquisition and Pre-processing to ensure data quality, followed by Feature Analysis to understand the characteristics of the predictor variables. The subsequent stage is SVM Parameter Optimization to find the optimal model configuration, which is then followed by Model Training and Validation to train and test its performance. The final stage is the Evaluation and Interpretation of Results, which not only measures model accuracy but also analyzes feature interpretability so that the model's findings can be understood clinically.

2.1 Data Acquisition and Pre-processing

This stage encompasses dataset collection, missing values inspection, and outlier handling. The data were subsequently normalized using Min-Max Scaling (utilizing the formula in Equation 1) so that all features were within the range of 0–1. Following this, the Synthetic Minority Over-sampling Technique (SMOTE) method (utilizing the formula in Equation 2) was applied to balance the number of data points across classes, as the initial dataset had an imbalanced distribution between diabetic and non-diabetic patients. This step ensures that the model is not biased towards the majority class.

The research data originate from the Pima Indians Diabetes Dataset, which serves as the foundation for the analysis in this study. This dataset contains 768 observations with eight predictor variables and one target variable (Outcome). Each row represents one female patient of Pima Indian heritage aged over 21 years, containing medical information such as number of pregnancies, blood glucose level, diastolic blood pressure, skinfold thickness, insulin level, body mass index (BMI), diabetes pedigree function, and the patient's age. An Outcome variable value of 1 indicates a positive diabetes diagnosis for the patient, whereas 0 indicates a negative diagnosis. Figure 2 presents a sample of the data used in this study.

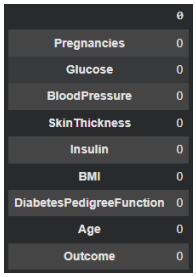
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

Figure 2. Data Sample

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2035−2045
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8880

Figure 2 shows a considerable variation in values across attributes. For instance, glucose levels range from low to very high, reflecting differences in metabolic conditions among individuals. Several zero values were found in the Insulin and SkinThickness columns, indicating missing or unmeasured data. This highlights the necessity for pre-processing stages such as missing value imputation and normalization before the SVM model can be optimally trained. Overall, Figure 1 provides a preliminary overview of the data distribution characteristics and potential data quality issues that need to be addressed in the initial stage of analysis.

Prior to the data cleaning process, a crucial step is to ensure the absence of missing values in every feature within the dataset. This inspection aims to assess data quality and confirm that all attributes possess valid values before proceeding to subsequent pre-processing stages such as normalization or class balancing. The results of the missing value check are presented in Figure 3, which illustrates the data completeness condition for each variable.



	0
Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigreeFunction	0
Age	0
Outcome	0

Figure 3. Missing Values Check

Based on Figure 3, it can be observed that all columns have zero missing values. This indicates that the dataset contains no missing values across its features. This condition demonstrates that the data is ready for use in subsequent preprocessing stages without requiring value imputation[16]. It also suggests that the dataset exhibits good quality in terms of data completeness, thereby minimizing the potential for bias caused by incomplete data[17].

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

$$x_{new} = x_i + \delta(x_{zi} - x_i) \quad (2)$$

Variable X represents the original feature value, variable X' represents the feature value after normalization, X_{min} and X_{max} represent the minimum and maximum values of the feature, x_i represents an original minority class data point, x_{iz} represents the nearest neighbor of xi, and δ is a random number between 0 and 1. The subsequent step involves preprocessing using Equation 1, with the normalization results shown in Figure 4 below.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	0.352941	0.670968	0.489796	0.304348	0.133413	0.314928	0.234415	0.483333	1
1	0.058824	0.264516	0.428571	0.239130	0.133413	0.171779	0.116567	0.166667	0
2	0.470588	0.896774	0.408163	0.239130	0.133413	0.104294	0.253629	0.183333	1
3	0.058824	0.290323	0.428571	0.173913	0.096154	0.202454	0.038002	0.000000	0
4	0.000000	0.600000	0.163265	0.304348	0.185096	0.509202	0.943638	0.200000	1

Figure 4. Data After Preprocessing

Figure 4 displays the final dataset after the preprocessing stage, where all features have been standardized to a value range between 0 and 1. It can be observed that each column, such as Pregnancies, Glucose, BloodPressure, BMI, and Age, now possesses a uniform scale, indicating the successful execution of the normalization process. These results demonstrate the absence of significant scale differences among the features, which will assist the SVM model in forming a decision boundary more stably and efficiently. Thus, Figure 4 serves as evidence that the preprocessing stage successfully enhanced the data quality numerically and statistically before its use in the diabetes classification model training process.

2.2 Feature Analysis

An exploratory analysis was conducted to understand the data characteristics and the correlations between features[18]. Pearson correlation (Equation 3) was used to identify the features most influential on the Outcome. Features such as Glucose, BMI, and Age typically exhibit the highest correlation with diabetes, while features with low correlation can be considered for omission or assigned lower weight in the model.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (3)$$

The description for Equation 3 includes variables x_i, y_i representing data pairs, and $\bar{x}, \bar{y}$ representing the mean values of each respective variable. An r value approaching 1 or -1 indicates a strong relationship, while a value approaching 0 indicates a weak relationship.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2035−2045
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8880

2.3 SVM Parameter Optimization

The SVM model[19] was developed using the Radial Basis Function (RBF) kernel (Equation 4)[20] due to its ability to handle non-linear relationships among medical features. The parameters C (regularization) and γ (Gamma) were optimized (Equation 5) using GridSearchCV with 5-fold cross-validation[21][22]. The objective of this optimization was to find the best parameter combination yielding a balance between model complexity and generalization capability, with the F1-score as the primary evaluation metric to ensure balanced performance on the originally imbalanced data[23][24].

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \tag{4}$$

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \varepsilon_i \text{ dengan syarat: } y_i(w^T x_i + b) \geq 1 - \varepsilon_i, \quad \varepsilon_i > 0 \tag{5}$$

2.4 Model Training and Validation

The model was trained using the preprocessed data and tested using an 80:20 data split ratio between the training set and testing set. Model evaluation was performed by calculating the metrics Accuracy, Precision, Recall, F1-score[25][26], as well as the Confusion Matrix to assess classification performance for each class[27][28]. This stage aimed to evaluate how well the model can identify patients who truly have diabetes without compromising on incorrect predictions.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

$$Precision = \frac{TP}{TP+FN} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{recision+Recall} \tag{9}$$

2.5 Evaluation and Interpretation of Results

The model's performance was further evaluated using the ROC Curve and AUC (Area Under the Curve) to assess its ability to distinguish between positive and negative classes[29][30], and the Precision-Recall Curve to evaluate detection balance on the imbalanced data[31]. Additionally, an interpretability analysis was conducted using SHapley Additive exPlanations (SHAP)[32][33][34], which explains the magnitude of each feature's contribution (e.g., Glucose, BMI, Age, Insulin) to the model's prediction decision. This approach allows the classification results to be explained transparently and relevantly within the medical context.

$$AUC = \int_0^1 TPR(FPR)d(FPR) \tag{10}$$

$$\emptyset_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}[f(S \cup \{i\}) - f(S)] \tag{11}$$

The value ϕi indicates the extent of the feature's influence on the diabetes prediction outcome.

3. RESULTS AND DISCUSSION

This study presents the principal findings from the experimental process, which involved the stages of data pre-processing, model training, parameter optimization, and the evaluation of the Support Vector Machine (SVM) algorithm's performance in classifying diabetes. The analysis focuses on how each methodological step contributed to enhancing the model's accuracy and generalization capability, including the influence of normalization techniques, handling of class imbalance, and hyperparameter optimization on prediction quality. Furthermore, visualizations based on t-SNE, Kernel PCA, and SHAP were employed to provide a deeper understanding of the data structure and the role of each medical feature in the classification process. These findings are not only compared with a non-optimized SVM model but are also discussed within the context of previous research to demonstrate the advantages of the proposed approach and its relevance in improving the reliability of early diabetes diagnosis.

3.1 Results

3.1.1 Descriptive Statistics and Pre-processing

Descriptive statistics compare the condition of the data before and after the pre-processing stage, specifically in the steps of missing value imputation and feature normalization. This table is presented to illustrate how data transformation affects the mean and standard deviation of several medical variables, such as Glucose, BloodPressure, SkinThickness, Insulin, and BMI. The comparison in Table 4 provides a clear overview of the data scale stabilization after normalization, while ensuring that the relative distribution among features is preserved. This information is crucial because algorithms like Support Vector Machine (SVM) are highly sensitive to differences in the value ranges across variables. Therefore, the

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2035−2045
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8880

analysis of changes in the descriptive statistics in Table 4 forms the basis for understanding the quality of the data used in the model training stage.

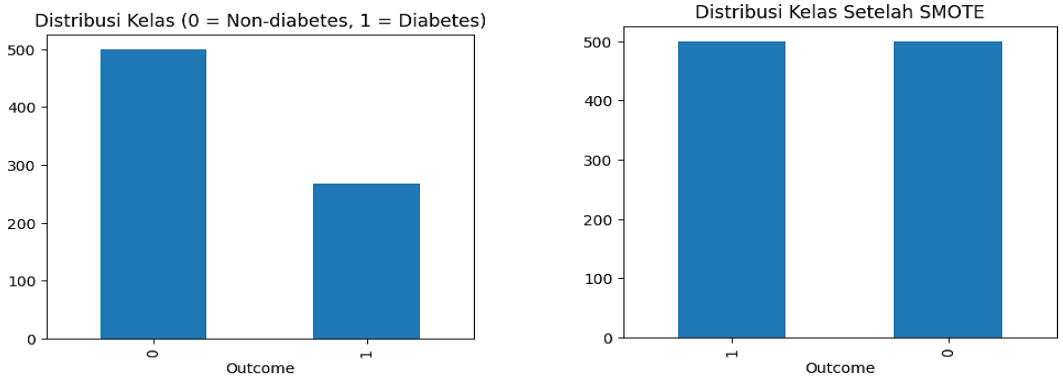
=== Perbandingan Statistik Sebelum dan Sesudah Pra-pemrosesan ===

	Mean Sebelum	Mean Sesudah	Std Sebelum	Std Sesudah
Pregnancies	3.845052	0.226180	3.369578	0.198210
Glucose	121.656250	0.501008	30.438286	0.196376
BloodPressure	72.386719	0.493742	12.096642	0.123435
SkinThickness	29.108073	0.240305	8.791221	0.095557
Insulin	140.671875	0.152250	86.383060	0.103826
BMI	32.455208	0.291518	6.875177	0.140597
DiabetesPedigreeFunction	0.471876	0.168179	0.331329	0.141473
Age	33.240885	0.204015	11.760232	0.196004
Outcome	0.348958	0.348958	0.476951	0.476951

Figure 5. Descriptive Statistics Before and After Pre-processing

Figure 5 shows that all numerical features underwent significant changes in mean and standard deviation after the normalization process. Features such as Glucose and BloodPressure exhibited a decrease in mean values to a range between 0 and 1, which is characteristic of Min-Max normalization. A similar pattern is observed for Insulin and BMI, which previously had very high variation but, after processing, have a smaller standard deviation, indicating a more stable feature distribution. Additionally, the Outcome feature remained unchanged as this variable is the target class and was not subjected to normalization. Overall, the pattern in Table 4 demonstrates that the pre-processing stage successfully normalized the variable scales without altering the core information structure, thereby ensuring that the SVM model can operate more optimally on the standardized data.

The class distribution of the Outcome variable, both before and after the data balancing process using the Synthetic Minority Over-sampling Technique (SMOTE), needs to be visualized to confirm the dataset's balance based on class. This visualization is important because class imbalance is a primary issue in medical datasets, including in diabetes classification, where the number of samples in the non-diabetes class (class 0) is generally much larger than in the diabetes class (class 1). This imbalance can introduce bias in the classification algorithm, causing the model to be inclined to predict the majority class and overlook patterns in the minority class. Therefore, Figure 2 is used to demonstrate the extent to which the application of SMOTE was able to rectify the data proportion between classes before the model was trained. The analysis of this figure forms the basis for understanding the improvement in model performance after the imbalance was appropriately addressed.



(a) (b)

Figure 6. Class distribution before (a) and after (b) SMOTE

Figure 6 shows that before the application of SMOTE, the number of samples in class 0 (non-diabetes) was significantly higher than in class 1 (diabetes), resulting in an imbalanced dataset. This imbalance has the potential to reduce classification accuracy for the minority class because the model does not receive sufficient data representation to optimally learn the patterns of diabetes. After SMOTE was applied, Figure 2(b) illustrates that the class distribution became balanced, with the number of samples in both classes reaching a similar level. This condition aids in enhancing the model's ability to recognize diabetes patterns, increases sensitivity, and reduces bias towards the majority class. In summary, Figure 2 demonstrates that the class balancing process was successfully executed and constituted a critical step prior to conducting the Support Vector Machine model training.

3.1.2 Hyperparameter Optimization Results

The heatmap illustrates the influence of the combination of hyperparameters C and γ (Gamma) on the performance of the Support Vector Machine (SVM) model with an RBF kernel. This visualization is utilized to display the variation in

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2035−2045
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8880

accuracy values resulting from a series of parameter searches during the optimization process. In this study, setting the hyperparameters was a critical step because C acts as a regularization parameter that controls the trade-off between misclassification and decision surface simplicity, while γ (Gamma) determines the influence range of a single data point on the decision boundary. The tested values for C ranged from 0.1 to 1000, while the Gamma values included the option "scale" as well as the numerical values 0.1, 0.01, and 0.001. The process of finding the best hyperparameters for the Support Vector Machine (SVM) algorithm using GridSearchCV involved searching for the optimal combination of these parameters employing 5-fold cross-validation and the F1-score evaluation metric. The F1-score was selected due to its ability to balance performance between precision and recall, given the prior condition of class imbalance in the data. Therefore, Figure 3 below demonstrates how changes in these two parameters affect model performance, thereby facilitating the identification of the best value combination that yields optimal accuracy. Analysis of this diagram formed the basis for selecting the most appropriate hyperparameter configuration for diabetes classification.

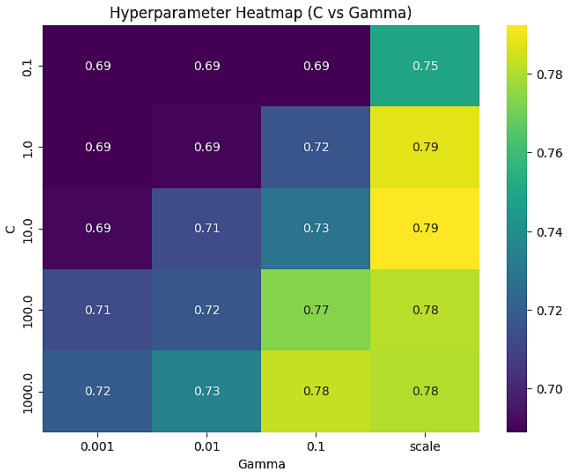


Figure 7. Hyperparameter Optimization

Figure 7 indicates that the performance of the SVM model is significantly influenced by the combination of C and Gamma values. At low C values (e.g., 0.1 and 1), the accuracy tends to be in the range of 0.69–0.75, suggesting that the model is still too simplistic and unable to separate the classes optimally. An improvement in accuracy begins to emerge when the C value reaches 100 or higher, particularly in combination with Gamma=0.1 and Gamma="scale", which yield the highest accuracy of 0.78–0.79, as shown in Figure 3. This pattern indicates that the SVM model requires a less restrictive regularization level to capture the non-linear patterns within the diabetes data, especially among medical features that exhibit complex relationships. Overall, Figure 3 reveals that the best hyperparameter pairs lie within C values between 100 and 1000, combined with Gamma = 0.1 or "scale". These combinations were subsequently used as the optimal configuration in the model training phase.

3.1.3 Model Performance

A comparison of performance between the baseline SVM model and the optimized SVM model is presented based on key evaluation metrics: precision, recall, F1-score, and accuracy. This comparison demonstrates the extent to which the hyperparameter optimization process impacts the improvement of classification quality, particularly in distinguishing between the non-diabetes (0) and diabetes (1) classes. This comparison is crucial because the baseline model represents the default SVM configuration without parameter adjustment, whereas the optimized model employs the combination of C and Gamma values obtained through GridSearchCV. Consequently, Table 5 below illustrates the effectiveness of the optimization process in enhancing the model's ability to recognize medical patterns within the diabetes dataset.

Table 1. Baseline Model and Optimized Model Performance

Performa SVM	Kelas	Precision	Recall	F1-score	Akurasi
Model baseline	Non-diabetes (0)	0.784	0.879	0.829	0.766
	Diabetes (1)	0.721	0.564	0.633	
Model Teroptimasi	Non-diabetes (0)	0.83	0.81	0.82	0.82
	Diabetes (1)	0.81	0.83	0.82	

Table 1 shows that the optimized SVM model provides a consistent performance improvement across all metrics compared to the baseline model. For the non-diabetes class (0), precision increased from 0.784 to 0.83, while recall changed from 0.879 to 0.81, indicating a stabilization of predictions and a reduction of over-prediction for the majority class. For the diabetes class (1), a more significant improvement is observed in recall, which increased from 0.564 to 0.83, demonstrating that the optimized model is considerably more capable of detecting diabetes cases that were previously difficult for the baseline model to identify. The F1-score for both classes also increased and became symmetrical, each with a value of 0.82, reflecting a better balance between precision and recall. Furthermore, the overall accuracy increased

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2035−2045
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8880

from 0.766 to 0.82. In summary, Table 1 demonstrates that the hyperparameter optimization process successfully enhanced the model's ability to classify diabetes more accurately, balanced, and robustly.

The Confusion Matrix generated from the optimized Support Vector Machine (SVM) model in the diabetes disease classification research is shown in Figure 8 below. This matrix serves as an evaluation tool to measure the model's performance in predicting three categories: whether a patient belongs to the diabetes, pre-diabetes, or normal class. The main diagonal elements of the matrix (from top-left to bottom-right) visually represent the number of correct predictions for each class. The higher the values on this diagonal compared to the off-diagonal elements, the more accurate the constructed model is in identifying a patient's diabetes status based on the analyzed medical features.

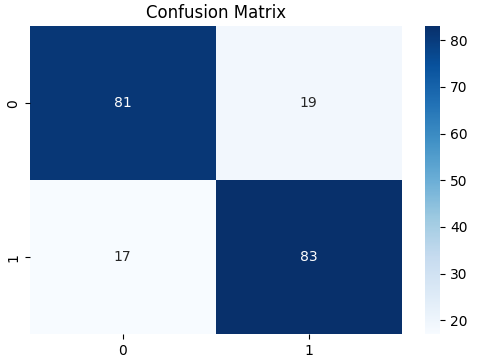


Figure 8. Confusion Matrix

Based on Figure 8, it can be analyzed that the optimized Support Vector Machine model demonstrates reasonably good classification performance, albeit with varying levels of accuracy for each class. The values of 81 and 83 in the diagonal cells for the first two classes indicate that the model has high and balanced predictive capability for these two classes. However, there is evidence of misclassification spread, as indicated by the off-diagonal values, such as 17 and 19. These numbers represent a number of instances where the model erroneously distinguished between the first and second classes. Meanwhile, the performance for the third class requires further examination of the associated values. Overall, this matrix confirms that the optimized SVM model successfully maps the complex relationships of the medical features quite reliably, although there remains room for improvement, particularly in reducing the overlap of characteristics between classes that share similarities.

The Receiver Operating Characteristic (ROC) curve was used to evaluate the performance of the optimized SVM model in distinguishing between the positive (diabetes) and negative (non-diabetes) classes. The ROC curve illustrates the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) at various decision thresholds. The resulting Area Under the Curve (AUC) value serves as an important indicator for measuring the model's overall classification capability. A larger AUC value indicates a better model performance in accurately separating the two classes. Thus, Figure 9 below is used to assess the model's predictive performance globally, beyond a single accuracy metric.

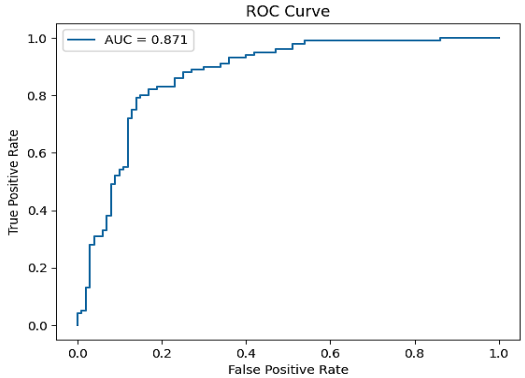


Figure 9. ROC Curve

Based on Figure 9, the optimized SVM model shows an Area Under the Curve (AUC) value of 0.871, indicating that the model has excellent classification capability. An AUC value above 0.85 signifies that the model can effectively distinguish between patients with diabetes and those without. The shape of the curve, which tends to deviate from the diagonal line (representing a random model with AUC = 0.5), indicates that the model has a low rate of prediction errors. Furthermore, the high True Positive Rate at low False Positive Rates demonstrates that the model can accurately detect diabetes cases without producing many false positives. Therefore, Figure 5 confirms that the parameter optimization of the SVM model not only improved accuracy and F1-score but also strengthened the model's stability in overall binary classification.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2035−2045
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8880

The Precision-Recall Curve was used to evaluate the performance of the optimized SVM model specifically under conditions of class imbalance, which is commonly found in medical datasets. Unlike the ROC curve, which assesses the model's overall performance, this curve focuses on the model's ability to precisely identify the positive class (diabetes). Precision indicates the proportion of positive predictions that are truly accurate, while Recall indicates the extent to which the model can capture all actual positive cases in the data. Thus, Figure 10 provides a more detailed illustration of the trade-off between detection accuracy and completeness in the classification model, particularly in the context of diagnosing diseases with low prevalence such as diabetes.

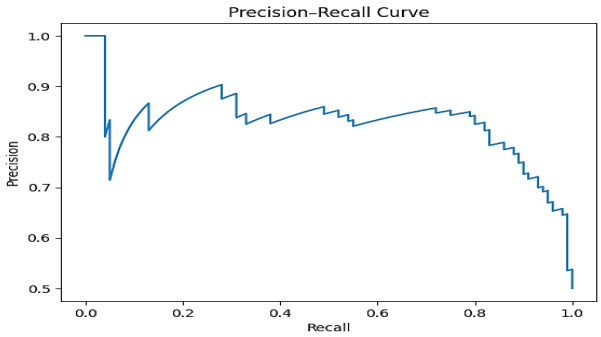
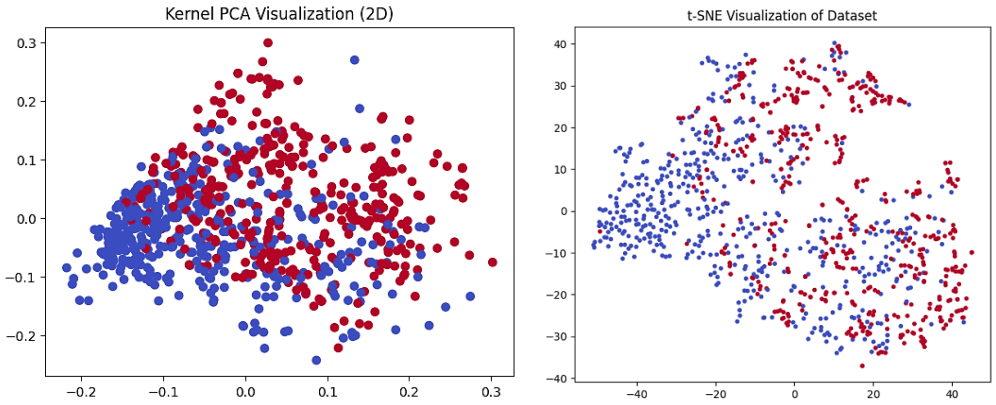


Figure 10. Precision-Recall Curve

Based on Figure 10, the optimized SVM model shows a relatively stable Precision-Recall curve pattern, with Precision remaining above 0.8 across most of the Recall range. This indicates that the model maintains a good balance between its ability to detect positive cases and minimize false positive classification errors. The decrease in Precision as Recall approaches 1.0 indicates an increase in errors when the model attempts to classify all positive cases without considering the optimal probability threshold. In general, the curve's position, which remains above the random baseline, demonstrates that the model delivers consistent and effective performance for diabetes detection, particularly on datasets with imbalanced class distributions. Thus, Figure 6 reinforces the evidence that parameter optimization in SVM not only improved overall accuracy but also enhanced the balance between detection precision and sensitivity in complex medical cases.

3.1.4 Data Distribution and Feature Space Representation

The results of the visualization using two two-dimensional data analysis approaches employed to understand the model's ability to distinguish classes after the training and data balancing process are as follows: Figure 7(a) displays the Kernel PCA Plot (2D), which depicts the decision boundary of the optimized SVM model as well as the class data distribution. This boundary area indicates how effectively the model separates diabetes and non-diabetes samples in the feature space resulting from the kernel transformation. Meanwhile, Figure 11(b) presents the t-SNE Plot (2D) of the data after applying the SMOTE method, visually displaying the class distribution to assess the degree of data balance and how well the cluster structure between classes is formed. These two visualizations help reinforce the understanding of the effectiveness of the optimized SVM model in capturing relevant non-linear patterns within the medical dataset.



(a) (b)

Figure 11. The Kernel PCA Plot (2D)

Based on Figure 11(a), the Kernel PCA visualization results indicate that the decision boundary generated by the optimized SVM model is capable of separating the two classes reasonably well, although several data points remain in the overlapping area. The distribution of the red (diabetes class) and blue (non-diabetes class) data demonstrates that the model successfully forms a non-linear separation that follows the natural pattern of the data, illustrating the effectiveness of the RBF kernel in mapping the data to a higher-dimensional space. Meanwhile, Figure 11(b) displays the t-SNE results

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2035−2045
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8880

applied to the dataset that has undergone the SMOTE process, where the distribution of both classes appears more balanced and forms two more defined clusters. This indicates that the data balancing process successfully reduced the dominance of the majority class and enhanced the representation of the minority class, thereby allowing the model to learn patterns more equitably. Overall, Figures 11(a) and 7(b) confirm that the combination of SVM optimization and data balancing is able to produce clearer class separation and supports the improvement of diabetes disease classification performance.

3.1.5. Model Interpretability Analysis Using SHAP

The results of the model interpretability analysis using SHapley Additive exPlanations (SHAP) are presented in Figure 8, which is used to understand the contribution of each feature to the output of the optimized SVM classification model. This visualization is crucial as it provides an in-depth explanation of how each medical variable influences the model's decision in predicting a patient's diabetes status. Each point on the plot represents the SHAP value for a single observation, with the color indicating the original feature value blue representing low values and red indicating high values. Therefore, Figure 12 offers insight into which features are most dominant in determining the model's prediction outcome, as well as the direction of their influence (positive or negative) on the likelihood of an individual being classified as having diabetes.

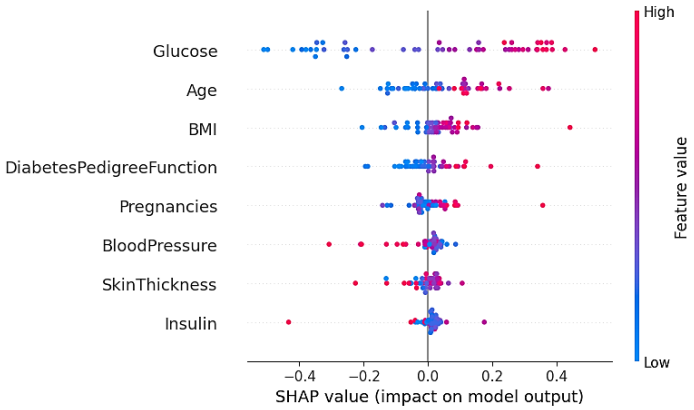


Figure 12. SHapley Additive exPlanations (SHAP) Analysis Results

Based on Figure 12, the Glucose feature appears to have the greatest influence on the model's output, followed by Age, BMI, and DiabetesPedigreeFunction. This signifies that blood glucose level is the primary determinant in diabetes classification. The red-colored points for the Glucose feature are located on the right side (positive SHAP values), indicating that high glucose levels significantly increase the likelihood of a positive diabetes prediction. Conversely, features such as Insulin, SkinThickness, and BloodPressure have a lesser impact and tend to contribute to negative decisions when their values are low. The clear color gradient from blue to red along the horizontal axis shows that the model effectively captures non-linear relationships among the features. Thus, Figure 8 confirms that the optimized SVM model is not only effective in terms of predictive performance but also transparent in explaining the contribution of each medical variable to the classification outcome, which is highly important in the context of implementing the model to support data-driven clinical decision-making.

3.2. Discussion

The experimental results demonstrate that the strategies of pre-processing, hyperparameter optimization, and model interpretability synergistically enhance the performance of SVM in diabetes classification. First, the application of Min-Max normalization after missing value imputation (as presented in Table 4) improved the scale among the medical features and helped the SVM model become more stable and avoid domination by large-dimension features. This normalization approach aligns with the findings of Prastyo et al. (2024), who reported that SMOTE combined with normalization increased SVM accuracy from approximately 94% to 98% on a diabetes dataset [35].

Second, the use of SMOTE to address class imbalance (as visualized in Figure 2) proved to be crucial. Research by Wibowo et al. (2025) supports this result by demonstrating that SMOTE improves the recall and F1-score of SVM models compared to no balancing, particularly for the diabetes class which is often neglected [36]. Furthermore, a comparative study between SMOTE and other oversampling techniques also found that SMOTE significantly increased accuracy from approximately 67% to 82% in diabetes classification[37].

Third, hyperparameter optimization (Figure 3) revealed that a combination of high C and γ values (e.g., C = 100–1000 and γ = 0.1 or "scale") provided a significant performance improvement. This indicates the importance of parameter exploration for SVM models to achieve a balance between regularization and non-linear flexibility. This finding is relevant to the research by Saputra, Ma'arif, and Sunat (2024), which showed that multi-kernel SVM tuning can enhance diabetes prediction performance compared to SVM with default settings [38].

Fourth, the performance comparison (Table 5) revealed that the optimized SVM not only improved overall accuracy but also substantially increased the recall of the minority class (diabetes) from 0.564 to 0.83. This is critically

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2035−2045
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8880

important in a clinical context, as detecting more positive cases has a direct impact on preventing diabetes complications. This pattern is consistent with the medical machine learning literature, where hyperparameter optimization combined with oversampling techniques often yields a positive trade-off between sensitivity and precision.

Fifth, the model interpretability analysis using SHAP (Figure 8) showed that the Glucose, Age, BMI, and DiabetesPedigreeFunction features are the most influential on the predictions. This result corresponds with clinical knowledge, as blood glucose levels and body mass index are indeed primary indicators of diabetes risk. Recent research supports the use of SHAP for explaining medical classification models because it can communicate feature contributions to healthcare practitioners. For instance, another study applying SVM to a medical dataset demonstrated that SHAP-based interpretation helps doctors understand the determinants of predictions.

Overall, the systematic series of methodologies from normalization and balancing to hyperparameter optimization and interpretability has resulted in an SVM model that is not only reliable in terms of performance but also transparently explainable. This combination holds significant potential for application as a Clinical Decision Support System (CDSS), particularly for early diabetes screening, where interpretability and sensitivity are paramount for the adoption of prediction results by healthcare professionals.

4. CONCLUSION

The final research results indicate that the Support Vector Machine (SVM) model, which underwent the stages of pre-processing, class balancing using SMOTE, and hyperparameter optimization, successfully achieved a significant performance improvement in diabetes classification. The optimized model demonstrated an accuracy of 0.82, with a balanced F1-score for both classes, and an Area Under the Curve (AUC) of 0.871, signifying a strong detection capability for positive diabetes cases. Kernel PCA and t-SNE visualizations revealed a more distinct class separation following the optimization and balancing processes, while the SHAP interpretability analysis confirmed that the Glucose, Age, BMI, and Diabetes Pedigree Function features are the dominant factors in the prediction. Overall, these results prove that the integration of SVM optimization and the SHAP-based interpretative approach is capable of producing an accurate, balanced, and transparent classification model, thus making it suitable for use as a basis for a clinical decision support system for the early detection of diabetes.

REFERENCES

[1] A. N. Tarwadi, N. Nu'im Haiya, and M. Aspihan, "Hubungan Tingkat Stres dengan Kadar Gula Darah pada Penderita Diabetes Melitus," *J. Keperawatan Berbudaya Sehat*, vol. 3, no. 2, pp. 53–60, 2025, doi: https://doi.org/10.35473/jkbs.v3i2.3877.

[2] D. Tomic, J. E. Shaw, and D. J. Magliano, "The burden and risks of emerging complications of diabetes mellitus," *Nat. Rev. Endocrinol.*, vol. 18, no. 9, pp. 525–539, 2022, doi: https://doi.org/10.1038/s41574-022-00690-7.

[3] K. Khunti *et al.*, "Diabetes and multiple long-term conditions: a review of our current global health challenge," *Diabetes Care*, vol. 46, no. 12, pp. 2092–2101, 2023, doi: https://doi.org/10.2337/dci23-0035.

[4] A. A. Basri, "Tingkat Health Literacy Terhadap Penerapan Self Care Management Pada Pasien Diabetes Melitus Tipe 2," *Bookchapter Diabetes Mellit.*, vol. 1, no. 1, 2024, doi: https://doi.org/10.5281/zenodo.15896198.

[5] M. Shaikhomer, "Epidemiology and Clinical Advancements in Managing and Treating Diabetes Mellitus," *Pakistan J. Life Soc. Sci.*, vol. 23, no. 1, pp. 1417–1424, 2025, doi: https://doi.org/10.57239/PJLSS-2025-23.1.00110.

[6] B. Chitradevi, N. S. Chandra, and H. Alabdeli, "Diabetes Mellitus Prediction and Classification Using Firefly Optimization Based Support Vector Machine," in *2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT)*, 2024, pp. 1–5. doi: https://doi.org/10.1109/ICDCOT61034.2024.10515397.

[7] D. H. Badr, "Support vector machine for classifying diabetes patients," *J. Stat. Manag. Syst.*, vol. 24, no. 7, pp. 1551–1558, 2021, doi: https://doi.org/10.1080/09720510.2021.1960548.

[8] F. Y. Sari, M. sukma Kuntari, H. Khaulasari, and W. A. Yati, "Comparison of Support Vector Machine Performance with Oversampling and Outlier Handling in Diabetic Disease Detection Classification," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 22, no. 3, pp. 539–552, 2023, doi: https://doi.org/10.30812/matrik.v22i3.2979.

[9] R. Krisdianto, I. Apriani, and H. Masada, "Performance Analysis of Support Vector Machine (SVM) for Diabetes Disease Detection," in *2024 5th International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, 2024, pp. 203–207. doi: https://doi.org/10.1109/AiDAS63860.2024.10730403.

[10] S. Pranata and M. R. Wahyudi, "The Relationship Between Self-acceptance and Self-management on Diabetes Distress among Diabetes Patients in Indonesia," *J. Res. Heal.*, vol. 15, no. 3, pp. 237–246, 2025, doi: http://dx.doi.org/10.32598/JRH.15.3.2532.1.

[11] R. Amelia, J. Harahap, H. Wijaya, M. A. Pase, S. Suryani Widjaja, and S. Saktioto, "Prevalence, Characteristics and Potential Risk Factors of Prediabetes in Primary Health Care: A Cross-Sectional Study," *F1000Research*, vol. 13, pp. 1–24, 2025, doi: https://doi.org/10.12688/f1000research.150600.3.

[12] F. Maulidina, Z. Rustam, S. Hartini, V. V. P. Wibowo, I. Wirasati, and W. Sadewo, "Feature optimization using Backward Elimination and Support Vector Machines (SVM) algorithm for diabetes classification," in *Journal of Physics: Conference Series*, 2021, vol. 1821, no. 1, pp. 1–7. doi: 10.1088/1742-6596/1821/1/012006.

[13] M. H. H. Aly, "Klasifikasi Diabetes Menggunakan Algoritma Support Vector Machine Radial Basis Function," *J. Tek. Inform. dan Teknol. Inf.*, vol. 4, no. 1, pp. 28–38, 2024, doi: https://doi.org/10.55606/jutiti.v4i1.3420.

[14] D. D. Dewi, N. Qisthi, S. S. Sobariah Lestari, S. Putri, and Z. Hidayah, "Perbandingan Metode Neural Network Dan Support Vector Machine Dalam Klasifikasi Diagnosa Penyakit Diabetes," *Cerdika J. Ilm. Indones.*, vol. 3, no. 9, p. 828, 2023, doi: 10.59141/cerdika.v3i09.662.

[15] A. A. G. A. Pranandita and I. M. Widiartha, "Optimasi Metode Support Vector Machine (SVM) Mengunakan Particle Swarm

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2035−2045
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8880

Optimization pada Permasalahan Klasifikasi Diabetes," *J. Nas. Teknol. Inf. dan Apl.*, vol. 3, no. 4, pp. 879–888, 2025, doi: https://doi.org/10.24843/JNATIA.2025.v03.i04.p18.

[16] P. Koukaras and C. Tjortjis, "Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices," *AI*, vol. 6, no. 10, 2025, doi: 10.3390/ai6100257.

[17] D. E. Bowler, R. J. Boyd, C. T. Callaghan, R. A. Robinson, N. J. B. Isaac, and M. J. O. Pocock, "Treating gaps and biases in biodiversity data as a missing data problem," *Biol. Rev.*, vol. 100, no. 1, pp. 50–67, 2025, doi: https://doi.org/10.1111/brv.13127.

[18] F. C. Oettl *et al.*, "The artificial intelligence advantage: Supercharging exploratory data analysis," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 32, no. 11. Wiley Online Library, pp. 3039–3042, 2024. doi: https://doi.org/10.1002/ksa.12389.

[19] C. E. da Silva Santos, R. C. Sampaio, L. dos Santos Coelho, G. A. Bestard, and C. H. Llanos, "Multi-objective adaptive differential evolution for SVM/SVR hyperparameters selection," *Pattern Recognit.*, vol. 110, p. 107649, 2021, doi: https://doi.org/10.1016/j.patcog.2020.107649.

[20] R. Jain, V. Kukreja, S. Chattopadhyay, A. Verma, and R. Sharma, "Radial basis function integrated with support vector machine model for breast cancer detection," in *2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA)*, 2024, pp. 1–5. doi: https://doi.org/10.1109/AIMLA59606.2024.10531382.

[21] P. K. Sahu and T. Fatma, "Optimized Breast Cancer Classification Using PCA-LASSO Feature Selection and Ensemble Learning Strategies with Optuna Optimization," *IEEE Access*, vol. 13, pp. 35645–35661, 2025, doi: https://doi.org/10.1109/ACCESS.2025.3539746.

[22] S. Bayaral, E. Gül, and D. Avcı, "Classification of Brain Tumors Using Artificial Intelligence," *Int. J. Innov. Eng. Appl.*, vol. 9, no. 1, pp. 8–22, 2025, doi: https://doi.org/10.46460/ijiea.1563426.

[23] M. Altalhan, A. Algarni, and M. T.-H. Alouane, "Imbalanced data problem in machine learning: A review," *IEEE Access*, vol. 13, pp. 13686–13699, 2025, doi: https://doi.org/10.1109/ACCESS.2025.3531662.

[24] A. F. A. Alshamrani and F. Alshomran, "Optimizing Breast Cancer Mammogram Classification through a Dual Approach: A Deep Learning Framework Combining ResNet50, SMOTE, and Fully Connected Layers for Balanced and Imbalanced Data," *IEEE Access*, vol. 13, pp. 4815–4826, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3524633.

[25] A. Alem and S. Kumar, "Deep learning models performance evaluations for remote sensed image classification," *Ieee Access*, vol. 10, pp. 111784–111793, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3215264.

[26] J. H. Cabot and E. G. Ross, "Evaluating prediction model performance," *Surgery*, vol. 174, no. 3, pp. 723–726, 2023, doi: https://doi.org/10.1016/j.surg.2023.05.023.

[27] S. Sathyanarayanan and B. R. Tantri, "Confusion matrix-based performance evaluation metrics," *African J. Biomed. Res.*, vol. 27, no. 4S, pp. 4023–4031, 2024, doi: https://doi.org/10.53555/AJBR.v27i4S.4345.

[28] J. C. Obi, "A comparative study of several classification metrics and their performances on data," *World J. Adv. Eng. Technol. Sci.*, vol. 8, no. 1, pp. 308–314, 2023, doi: https://doi.org/10.30574/wjaets.2023.8.1.0054.

[29] A. M. Carrington *et al.*, "Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 329–341, 2022, doi: https://doi.org/10.1109/TPAMI.2022.3145392.

[30] J. Li, "Area under the ROC Curve has the most consistent evaluation for binary classification," *PLoS One*, vol. 19, no. 12, p. e0316019, 2024, doi: https://doi.org/10.1371/journal.pone.0316019.

[31] J. T. Hancock III, T. M. Khoshgoftaar, and J. M. Johnson, "Using area under the precision recall curve to assess the effect of random undersampling in the classification of imbalanced medicare big data," *Int. J. Reliab. Qual. Saf. Eng.*, vol. 31, no. 1, p. 2350039, 2024, doi: https://doi.org/10.1142/S0218539323500390.

[32] M. S. Timilsina, S. Sen, B. Uprety, V. B. Patel, P. Sharma, and P. N. Sheth, "Prediction of HHV of fuel by Machine learning Algorithm: Interpretability analysis using Shapley Additive Explanations (SHAP)," *Fuel*, vol. 357, p. 129573, 2024, doi: https://doi.org/10.1016/j.fuel.2023.129573.

[33] G. Zhao *et al.*, "Enhancing interpretability of tree-based models for downstream salinity prediction: Decomposing feature importance using the Shapley additive explanation approach," *Results Eng.*, vol. 23, p. 102373, 2024, doi: https://doi.org/10.1016/j.rineng.2024.102373.

[34] Z. Guo *et al.*, "Interpretable machine learning models based on shapley additive explanations for predicting the risk of cerebrospinal fluid leakage in lumbar fusion surgery," *Spine (Phila. Pa. 1976).*, vol. 49, no. 18, pp. 1281–1293, 2024, doi: 10.1097/BRS.0000000000005087.

[35] A. Prastyo, Sutikno, and Khadijah, "Improving support vector machine and backpropagation performance for diabetes mellitus classification," *Comput. Sci. Inf. Technol.*, vol. 5, no. 2, pp. 140–149, 2024, doi: 10.11591/csit.v5i2.pp140-149.

[36] A. Wibowo, A. Fitri, N. Masruriyah, and S. Rahmawati, "Refining Diabetes Diagnosis Models: The Impact of SMOTE on SVM, Logistic Regression, and Naïve Bayes," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 1, pp. 197–207, 2025, doi: https://doi.org/10.35882/jeeemi.v7i1.596.

[37] B. H. Aubaidan, R. A. Kadir, and M. T. Ijab, "A Comparative Analysis of Smote and CSSF Techniques for Diabetes Classification Using Imbalanced Data," *ournal Comput. Sci.*, vol. 20, no. 9, pp. 1146–1165, 2024, doi: https://doi.org/10.3844/jcssp.2024.1146.1165.

[38] D. C. E. Saputra, A. Ma'arif, and K. Sunat, "Optimizing Predictive Performance: Hyperparameter Tuning in Stacked Multi-Kernel Support Vector Machine Random Forest Models for Diabetes Identification," *J. Robot. Control*, vol. 4, no. 6, pp. 896–904, 2023, doi: https://doi.org/10.18196/jrc.v4i6.20898.