JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i2.6878

OPTIMIZATION OF SVM ALGORITHM FOR OBESITY CLASSIFICATION WITH SMOTE TECHNIQUE AND HYPERPARAMETER TUNING

Khairun Nisa Arifin Nur[1*]; Anjar Wanto[1]; Poningsih[1]

Informatics Study Program[1]
STIKOM Tunas Bangsa, Pematangsiantar, Indonesia[1]
www.stikomtunasbangsa.ac.id[1]
khairunnisa@amiktunasbangsa.ac.id*, anjarwanto@amiktunasbangsa.ac.id,
poningsih@amiktunasbangsa.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Excessive fat accumulation that impairs personal health and raises the risk of chronic diseases is the hallmark of obesity, a global health issue. Decision Tree (DT) has been widely used for obesity classification, but it tends to suffer from overfitting and poor performance on imbalanced datasets. To overcome these limitations, this study proposes an optimization of the Support Vector Machine (SVM) algorithm using Synthetic Minority Over-sampling Technique (SMOTE) and Hyperparameter Tuning. SMOTE was applied to balance the class distribution, whereas Grid Search was utilized to determine the optimal combination of hyperparameters (C, gamma, and kernel). The dataset employed in this research comprises multiple features related to individual health and lifestyle, with obesity level as the target class. The experimental results demonstrate that the optimized SVM model demonstrated strong classification performance, attaining 97% in accuracy, precision, recall, and F1-score. This high performance is significant because it enables more accurate early detection of obesity risk, which can support timely medical intervention and personalized treatment planning, ultimately contributing to better public health outcomesThese findings suggest that incorporating SMOTE and Hyperparameter Tuning substantially improves SVM performance, establishing it as a robust approach for obesity classification and early risk detection.*

Keywords*: classification, hyperparameter tuning, machine learning, obesity, SVM.*

Intisari— *Obesitas merupakan masalah kesehatan global yang ditandai dengan akumulasi lemak berlebih dan berdampak buruk terhadap kesehatan serta meningkatkan risiko penyakit kronis. Meskipun algoritma Decision Tree (DT) telah banyak digunakan untuk klasifikasi obesitas, namun DT memiliki keterbatasan seperti overfitting dan performa yang rendah pada data tidak seimbang. Penelitian ini bertujuan mengoptimalkan algoritma Support Vector Machine (SVM) dengan menerapkan teknik Synthetic Minority Over-sampling Technique (SMOTE) dan Hyperparameter Tuning untuk mengatasi masalah tersebut. SMOTE digunakan untuk menyeimbangkan distribusi kelas, sedangkan Grid Search digunakan untuk menentukan kombinasi parameter terbaik (C, gamma, dan kernel). Dataset yang digunakan berisi fitur-fitur terkait kesehatan dan gaya hidup individu dengan kelas target tingkat obesitas. Hasil eksperimen menunjukkan bahwa model SVM yang telah dioptimasi berhasil mencapai akurasi, precision, recall, dan f1-score sebesar 97%. Pencapaian ini memiliki makna penting karena memungkinkan deteksi dini risiko obesitas secara lebih akurat, sehingga dapat mendukung intervensi medis yang tepat waktu dan perencanaan perawatan yang dipersonalisasi, serta berkontribusi pada peningkatan kualitas layanan kesehatan masyarakat. Temuan ini menunjukkan bahwa integrasi SMOTE dan Hyperparameter Tuning secara signifikan meningkatkan performa SVM, menjadikannya pendekatan yang kuat untuk klasifikasi obesitas dan deteksi risiko sejak dini.*

Kata Kunci*: klasifikasi, machine learning, obesitas, SVM, tuning parameter.*

INTRODUCTION

Obesity is a significant health disorder marked by the abnormal accumulation of excessive body fat. [1], which negatively impacts overall health [2]. As defined by the WHO, obesity is described as a condition where an individual possesses a Body Mass Index (BMI) of 30 or above, calculated by dividing a person's body weight in kilograms by the square of their height measured in meters [3]. Obesity has become a significant global health concern, with its prevalence steadily rising across both developed and developing nations. [4], posing significant physical, social, and economic consequences [5]. As the number of individuals affected by obesity continues to rise, the demand for accurate detection and classification methods becomes increasingly urgent [6].

Various studies have been developed to build obesity prediction models based on machine learning. Among them is a study that proposed the use of CatBoost for obesity classification and achieved high accuracy [7], [8]. Another study compared several algorithms, such as K-Nearest Neighbors, Naïve Bayes, SVM, and Decision Tree, to evaluate the effectiveness of each approach [9]. Additionally, a study developed a data-driven classification approach to support healthcare practitioners in precisely determining the degree of obesity [10][11]. Another research integrated electronic health record data with machine learning techniques and emphasized the crucial role of artificial intelligence in detecting obesity and its variants [12], [13], [14].

Meanwhile, another study utilized the Decision Tree (DT) algorithm for predicting obesity levels and achieved an accuracy of 90%, comparable to that of neural networks, while highlighting the significant contribution of genetic factors in obesity classification [15]. Although DT demonstrates good performance, this model has limitations in handling imbalanced data [16], which often leads to bias toward the majority class and reduced sensitivity to minority classes [17]. Moreover, improper parameter selection in DT can result in overfitting or underfitting [18], as well as model instability due to small changes in the input data [19]. As an alternative, Support Vector Machine (SVM) provides advantages in minimizing overfitting and managing model complexity [20]. Support Vector Machine (SVM) functions by optimizing the separation margin between classes and demonstrates strong performance in managing high-dimensional datasets [21]. Nevertheless, SVM also struggles with imbalanced data, often prioritizing the majority class [22]. Its performance heavily relies on selecting the right hyperparameters such as C, gamma, and the kernel type [8], [23].

To enhance SVM performance, two complementary strategies can be adopted: Synthetic Minority Oversampling Technique (SMOTE) and Hyperparameter Tuning. The SMOTE technique produces additional synthetic data points for underrepresented classes, thereby improving class distribution and fostering fairness in the model [24], while Hyperparameter Tuning facilitates the search for the optimal parameter combination using methods such as Grid Search [25], [26]. Together, these techniques help the SVM model generalize better, reduce overfitting risks, and enhance classification accuracy, particularly for minority classes [27], [28].

Based on the aforementioned background, this study seeks to optimize the SVM algorithm for obesity classification by integrating SMOTE and Hyperparameter Tuning techniques. This method is anticipated to yield a more precise, balanced, and computationally efficient model, supporting early detection and informed decision-making in health-oriented machine learning applications.

The novelty of this study lies in the integrated use of SMOTE to handle class imbalance and Grid Search-based hyperparameter tuning to enhance model performance—both applied specifically to the SVM algorithm for obesity classification. In contrast to earlier research that predominantly emphasized a single aspect, this research combines both techniques in a unified framework to achieve a more robust and generalizable model for healthcare analytics.

MATERIALS AND METHODS

The approach employed in this research utilizes the SVM algorithm to classify obesity levels, enhanced through the application of the SMOTE and Hyperparameter Tuning. The research consists of five stages, namely:

Data Collection

The dataset used in this research was acquired through the Kaggle repository and comprises information pertaining to obesity levels. The dataset includes various input variables such as weight, height, gender, age, family history of overweight, frequency of physical activity, dietary habits, and other lifestyle-related attributes. The target variable, labeled "NObeyesdad", consists of seven classes: Overweight_Level_I, Overweight_Level_II, Insufficient_Weight, Obesity_Type_I, Obesity_Type_II, Obesity_Type_III

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i2.6878

and Normal_Weight. The data consists of 2111 records, with a balanced distribution across several classes but notable imbalance in others. Accordingly, the dataset was subjected to preprocessing procedures aimed at cleaning and transforming its features, followed by a balancing stage using SMOTE to ensure equal representation of all classes before training and evaluating the classification model.

The dataset is openly accessible on Kaggle and contains no personally identifiable information. Its utilization in this study adheres to the usage terms and licensing conditions specified by the dataset provider.

Data Processing

The dataset employed in this research comprises more than 2,000 individual records with a total of 17 variables, including both input features and one target variable. The data was obtained from the Kaggle platform and is specifically designed for the categorization of obesity levels according to demographic, lifestyle, and behavioral attributes. The dependent variable in this study is NObeyesdad, which denotes the obesity level classification for each individual.

The outcome variable examined in this study are height, gender, weight, and age, along with other behavioral and lifestyle-related variables, such as food consumption habits, smoking behavior, physical activity, alcohol intake, and use of technology devices. A comprehensive list of the variables along with their descriptions is provided in Table 1.

Table 1. Variables and Descriptions

No	Variable	Type	Description
1	Gender	Object	Individual's gender
2	Age	Float64	Individual's age
3	Height	Float64	The individual's height expressed in meters
4	Weight	Float64	The individual's weight expressed in kilograms
5	Family history with overweight	Object	A family history of obesity is present.
6	FAVC	Object	Frequency of high-calorie food intake
7	FCVC	Float64	Frequency of vegetable consumption
8	NCP	Float64	The number of main meals consumed per day
9	CAEC	Object	Snacking frequency between meals
10	SMOKE	Object	Does the person smoke?
11	CH2O	Float64	Daily water intake
12	SCC	Object	Monitoring of caloric intake
13	FAF	Float64	The number of times physical activity is performed per week
14	TUE	Float64	Time using technology devices in hours
15	CALC	Object	Consumption of alcohol
16	MTRANS	Object	Transportation used
17	NObeyesdad	Object	Obesity Level

Source : (Research Results, 2025)

Table 1 presents data sourced from Kaggle (https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster), where the data has already been categorized by type. Prior to modeling, the dataset underwent several preprocessing steps. First, missing or inconsistent data entries were checked and cleaned. Next, categorical variables, including MTRANS, SMOKE, Family history of overweight, SCC, CAEC, CALC, and Gender, were encoded into numerical format using label encoding where appropriate. Numerical attributes such as Age, Height, and Weight were subjected to standard scaling for normalization to ensure uniform value distribution and prevent bias during model training, as StandardScaler transforms features to have zero mean and unit variance, which helps improve SVM performance since it is sensitive to feature scales [29].

The list of datasets for each variable can be seen in table 2 below.

Table 2. Dataset

Variable / Sample	Data 1	2	3	...	2111
Gender	Fem	Fem	Male	...	Fem
Age	21	21	23	...	23.6
Height	1.62	1.52	1.8	...	1.7
Weight	64	56	77	...	133.4
family_history_with_overweight	Yes	Yes	Yes	...	Yes
FAVC	No	No	No	...	Yes
FCVC	2	3	2	...	3
NCP	3	3	3	...	3
CAEC	Sometimes	Sometimes	Sometimes	...	Sometimes
SMOKE	no	yes	no	...	No
CH2O	2	3	2	...	2.8
SCC	no	yes	No	...	No
FAF	0	3	2	...	1.02
TUE	1	0	1	...	0.7
CALC	no	Sometimes	Frequently	...	Sometimes
MTRANS	Public	Public	Public	...	Public
NObeyesdad	Normal_Weight	Normal_Weight	Normal_Weight	...	Obesity_TypeIII

Source : (Research Results, 2025)

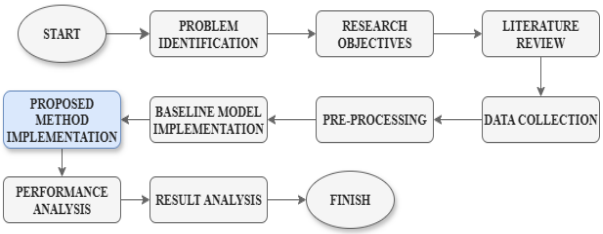
This data processing stage aims to ensure that the dataset used in the study is clean, balanced, and ready to be used in the classification process using the SVM algorithm.

VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i2.6878

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

Research Stages

This research was undertaken by following a structured series of methodological stages aimed at achieving the main objective, namely the development and optimization of an obesity classification model based on Support Vector Machine (SVM). Each stage was carefully designed to ensure an efficient workflow, starting from problem identification to model performance evaluation. Figure 1 illustrates the flowchart of the research stages implemented in this study.



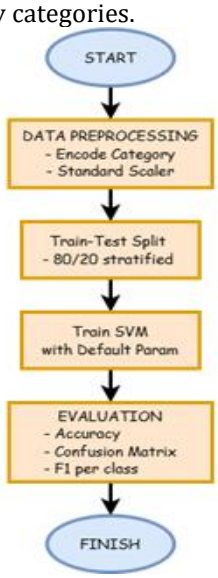
Source : (Research Results, 2025)

Figure 1. Research Framework

1. Problem Identification
   The process begins with identifying the core issues surrounding obesity classification, particularly its importance in disease prevention and data-driven decision-making.
2. Research Objectives
   Based on the identified problems, the main research objective is formulated: to build an accurate and optimized multiclass obesity classification model.
3. Literature Review
   A thorough literature review is conducted to understand relevant methods, including classification techniques, handling imbalanced data, and performance evaluation metrics. References are drawn from journals, books, and other reputable sources.
4. Data Collection
   The dataset applied in this study was retrieved from an open-source repository and encompasses attributes associated with individuals' lifestyle, dietary patterns, and physical conditions.
5. Pre-processing
   This stage encompasses encoding categorical features, standardizing the data, and segregating the dataset into distinct training and testing subsets. Such preprocessing steps guarantee that the dataset is properly structured for the training phase of the model.
6. Baseline Model Implementation
   A baseline SVM model is implemented as a benchmark before applying optimization techniques. The model is developed using the preprocessed dataset and assessed through performance metrics, including ROC curve, classification report, and the confusion matrix.
7. Proposed Method Implementation
   At this stage, the baseline model is enhanced by incorporating SMOTE for balancing the dataset and GridSearchCV for hyperparameter tuning. The optimized model is expected to perform better than the baseline.
8. Performance Analysis
   The optimized model is evaluated using the same performance metrics to observe improvements in accuracy and generalization. Class-wise F1-scores, confusion matrix, and ROC curves are analyzed and compared to the baseline.
9. Result Analysis
   To learn more about the model's advantages, disadvantages, and capacity to categorize different forms of obesity, the evaluation findings are further examined. The performance disparities between the baseline and optimized models are also emphasized.

Support Vector Machine (SVM) Algorithm

This research uses the SVM algorithm to classify obesity levels based on behavioral and physical attributes. SVM effectively handles high-dimensional data by creating decision boundaries that maximize class separation [30], [31]. To enhance its performance, this research utilizes the SMOTE to mitigate class imbalance and implements Hyperparameter Tuning to identify the optimal model parameters. The integration of these strategies are designed to improve the accuracy and generalization capability of SVM in classifying different obesity categories.



Source : (Research Results, 2025)

Figure 2. Support Vector Machine (SVM) Baseline

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i2.6878

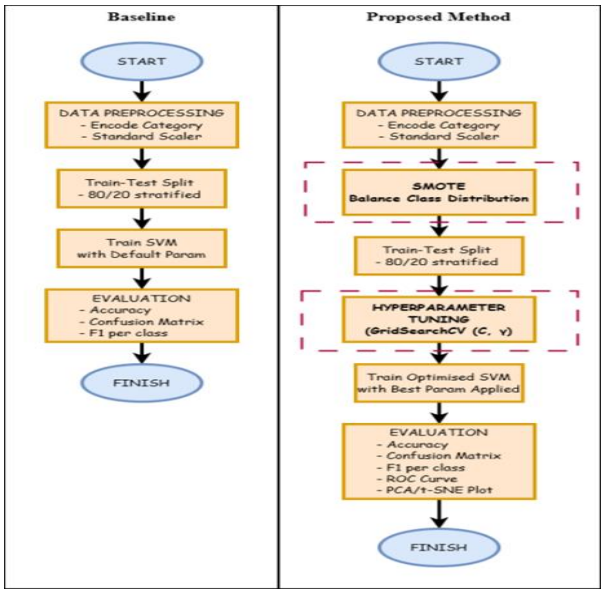
Proposed Method

To assess the effectiveness of the proposed approach in this study, a comparison was carried out between the baseline model and the optimized model. The baseline model employs a SVM classifier with default parameters, without addressing data imbalance or performing any parameter tuning. In contrast, the proposed model integrates two key optimization techniques: SMOTE to address class imbalance, and Hyperparameter Tuning using GridSearchCV to determine the best combination of hyperparameters (C and γ).

GridSearchCV was selected in this study due to its exhaustive and systematic nature in exploring the parameter space [32]. Although methods such as RandomSearchCV and Bayesian Optimization are known for their efficiency in larger and more complex hyperparameter spaces, GridSearchCV provides a more deterministic approach, ensuring all parameter combinations within the defined grid are tested [33]. Given the relatively small size of the hyperparameter space used in this study (C and γ), GridSearchCV was considered adequate and effective for achieving reliable optimization results without significantly increasing computational cost.

The parameter grid explored in this study includes the following values: C = [0.1, 1, 10], gamma = ['scale', 0.01, 0.001], and kernel = ['rbf', 'linear']. These combinations were evaluated using 5-fold cross-validation within GridSearchCV to determine the optimal SVM configuration.

The flowchart comparing the SVM Baseline with the Proposed Method is shown in Figure 3.



Source : (Research Results, 2025)

Figure 3. Comparison of SVM Baseline with Proposed Method

Figure 3 illustrates the step-by-step comparison of both approaches. On the left side, the baseline workflow begins with data preprocessing, which includes category encoding and standard scaling, followed by an 80:20 stratified train-test split. Subsequently, the model is trained with the default SVM configuration and evaluated using performance indicators, including confusion matrix, accuracy, and class-wise F1-score. On the right side, the proposed method workflow also starts with preprocessing, but continues with the application of SMOTE to balance the class distribution. After that, the data is split with the same stratified ratio (80:20), and hyperparameter tuning is carried out to find the optimal values for C and γ using GridSearchCV. The optimized SVM is then trained with the best parameters and evaluated with a more comprehensive set of metrics, including accuracy, confusion matrix, F1-score per class, ROC curve, and PCA/t-SNE visualization to observe the class separation visually.

This comparison clearly demonstrates the impact of applying SMOTE and hyperparameter optimization on the performance of obesity classification models, showing improvements in accuracy and class-level sensitivity.

RESULTS AND DISCUSSION

This section reports the outcomes of the classification experiments and elaborates on the performance comparison between the baseline SVM model and the optimized version enhanced with SMOTE and hyperparameter tuning. The evaluation emphasizes key metrics such as precision, accuracy, F1-score, recall, and the confusion matrix to comprehensively assess classification performance across all obesity categories. Additional visualizations such as ROC curves and dimensionality reduction plots (PCA/t-SNE) are also provided to support the interpretation of the model's behavior. The discussion further highlights the improvements gained through the application of oversampling and parameter optimization, and their impact on handling class imbalance and enhancing model generalization.

Data Pre-Processing Results

Prior to training the classification model, the dataset was subjected to a series of preprocessing steps to guarantee that all features were appropriately formatted for machine learning. This involved transforming categorical variables into numerical values and standardizing numerical features to ensure consistency in scale and prevent bias during model training. The results of these

preprocessing steps are presented in Table 3 and Table 4.

Table 3 shows the transformation of several categorical variables into numerical format using label encoding, where each category is assigned a unique numeric code. This step was necessary for algorithms such as SVM that require numeric input.

Table 3. Encoded Categorical Variables

Variable Sample	Data				
	1	2	3	...	2111
Gender	0	0	1	...	0
family_history_with_overweight	0	0	0	...	0
FAVC	0	0	0	...	1
CAEC	0	0	0	...	0
SMOKE	0	1	0	...	0
SCC	0	1	0	...	0
CALC	2	1	0	...	0
MTRANS	0	0	0	...	0
NObeyesdad	0	0	0	...	1

Source : (Research Results, 2025)

Table 4 presents the results of applying standard scaling to the numerical features in the dataset, including variables such as age, height, weight, and physical activity. StandardScaler transforms feature values such that they attain a mean of 0 and a standard deviation of 1, which helps stabilize the training process and improves the convergence of the SVM algorithm.

Table 4. Standard Scaled Numerical Variables

Variable Sample	Data				
	1	2	3	...	2111
Age	-0.980293	-0.980293	0.700914	...	1.259671
Height	-0.460357	-1.386461	1.206630	...	0.640188
Weight	-0.614415	-0.878422	-0.185404	...	1.678241
FCVC	-1.0	1.0	-1.0	...	1.0
NCP	0.0	0.0	0.0	...	0.0
CH2O	-0.994678	1.140382	-0.994678	...	0.848974
FAF	-1.351063	1.339203	0.442447	...	-0.430587
TUE	0.786431	-1.659960	0.786431	...	0.087098

Source : (Research Results, 2025)

The robustness and generalization of the SVM model are enhanced by these preprocessing outcomes, which guarantee that all data utilized for training and testing is in a consistent, numerical format and suitably scaled.

SMOTE and Pre-Processing

SMOTE was implemented to overcome the problem of class imbalance within the dataset. Prior to the implementation of SMOTE, The training data displayed a considerable imbalance between the majority and minority classes. Therefore, this subsection presents a comparison of the class distribution before and after applying SMOTE to demonstrate the effectiveness of this technique in balancing the number of instances across all classes. Class distribution before using the SMOTE technique is shown in Table 5.

Table 5. Amount of Data in Each Classes

No	Class	Amount
1	Insufficient_Weight	272
2	Normal_Weight	287
3	Overweight_Level_I	290
4	Overweight_Level_II	290
5	Obesity_Type_I	351
6	Obesity_Type_II	297
7	Obesity_Type_III	324

Source : (Research Results, 2025)

Table 5 presents the distribution of data in each class before the application of SMOTE. It can be observed that the dataset was imbalanced, with the number of samples varying significantly across classes. For example, *Obesity_Type_I* had the highest number of instances (351), while *Insufficient_Weight* had the fewest (272). This imbalance may adversely impact the performance of classification models, particularly in predicting minority classes.

Table 6. Amount of Data in Each Classes with SMOTE

No	Class	Amount
1	Insufficient_Weight	281
2	Normal_Weight	281
3	Overweight_Level_I	281
4	Overweight_Level_II	281
5	Obesity_Type_I	281
6	Obesity_Type_II	281
7	Obesity_Type_III	281

Source : (Research Results, 2025)

To resolve this problem, the SMOTE technique was implemented, with the resulting class distribution presented in Table 6. After the application of SMOTE, all classes were resampled to have an equal number of instances, specifically 281 samples per class. The goal of this balancing procedure is to improve the model's generalization across all categories by preventing the classifier from becoming biased toward the majority classes.

Data Classification Results

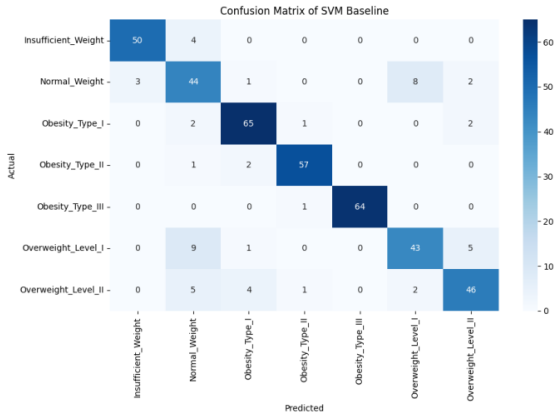
At this stage, a classification process was conducted to categorize individuals into one of the seven obesity levels based on various physical, demographic, and behavioral attributes. The classification was performed using two different approaches: the baseline SVM model with default settings and no data balancing, and the proposed SVM model optimized using the SMOTE technique for class balancing and Hyperparameter Tuning via GridSearchCV.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i2.6878

Confusion Matrix

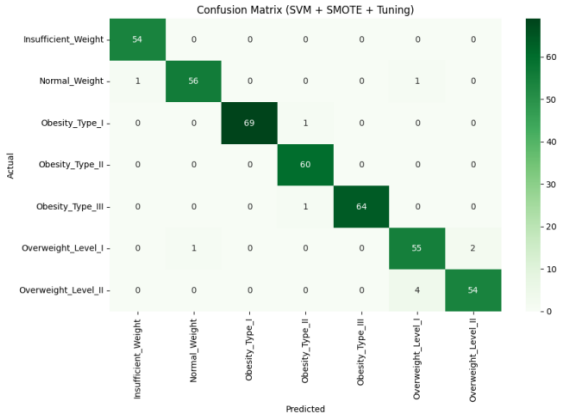
To further assess the performance of the classification models, confusion matrices were utilized to illustrate the accuracy in classifying the seven obesity level categories. Figure 4 presents the confusion matrix of the baseline SVM model, while Figure 5 illustrates the confusion matrix of the optimized SVM model enhanced with SMOTE and Hyperparameter Tuning.



Source : (Research Results, 2025)

Figure 4. Confusion Matrix SVM Baseline

In the Figure 4 baseline SVM model, misclassifications occurred in several classes, especially in Overweight_Level_I, Overweight_Level_II, and Normal_Weight. For example, Overweight_Level_I was misclassified into Normal_Weight in 9 instances and into Overweight_Level_II in 5 instances. Similarly, Normal_Weight showed misclassification into Insufficient_Weight and Overweight_Level_I. These errors highlight the model's limited sensitivity to minority classes due to the imbalanced distribution of training data.



Source : (Research Results, 2025)

Figure 5. Confusion Matrix SVM Proposed

In contrast, in Figure 5 the proposed SVM model with SMOTE and Hyperparameter Tuning shows significant performance improvement across all categories. The confusion matrix shows more accurate classification, with fewer misclassifications. Most categories, such as Insufficient_Weight, Obesity_Type_I, Obesity_Type_II, and Obesity_Type_III achieve near-perfect or perfect predictions. The SMOTE implementation successfully addresses the class imbalance problem, while GridSearchCV improves the model generalization by finding optimal C and γ values.

The confusion matrix clearly indicates that, following the application of SMOTE and hyperparameter tuning, the number of misclassified instances in minority classes (such as *Normal_Weight* and *Overweight_Level_I*) was significantly reduced. This indicates that the optimized model exhibits improved capability in identifying previously underrepresented categories.

Overall, the proposed method proved effective in improving classification accuracy and reducing class prediction bias, particularly for the minority categories. This improvement is reflected in the significant increase in evaluation metrics such as precision, recall, and F1-score in the next sub-section.

Classification Report

This section provides a comparative evaluation of classification performance between the baseline SVM model and the proposed approach, which combines SVM with SMOTE for class balancing and hyperparameter tuning for optimization.

Classification Report:

	precision	recall	f1-score	support
Insufficient_Weight	0.94	0.93	0.93	54.00
Normal_Weight	0.68	0.76	0.72	58.00
Obesity_Type_I	0.89	0.93	0.91	70.00
Obesity_Type_II	0.95	0.95	0.95	60.00
Obesity_Type_III	1.00	0.98	0.99	65.00
Overweight_Level_I	0.81	0.74	0.77	58.00
Overweight_Level_II	0.84	0.79	0.81	58.00
accuracy	0.87	0.87	0.87	0.87
macro avg	0.87	0.87	0.87	423.00
weighted avg	0.87	0.87	0.87	423.00

Source : (Research Results, 2025)

Figure 6. Classification Report SVM Baseline

As illustrated in Figure 6, the baseline SVM model attained an overall accuracy of 87%, with both the macro and weighted average F1-scores reaching 0.87. The model demonstrated strong performance on classes like Obesity_Type_II and Obesity_Type_III, achieving F1-scores of 0.95 and 0.99, respectively. However, the model's performance declined notably for Overweight_Level_I and Normal_Weight, with F1-

scores of 0.72 and 0.77, highlighting its challenges in differentiating among intermediate obesity levels.

Classification Report (SVM + SMOTE + Tuning):

	precision	recall	f1-score	support
Insufficient_Weight	0.98	1.00	0.99	54.00
Normal_Weight	0.98	0.97	0.97	58.00
Obesity_Type_I	1.00	0.99	0.99	70.00
Obesity_Type_II	0.97	1.00	0.98	60.00
Obesity_Type_III	1.00	0.98	0.99	65.00
Overweight_Level_I	0.92	0.95	0.93	58.00
Overweight_Level_II	0.96	0.93	0.95	58.00
accuracy	0.97	0.97	0.97	0.97
macro avg	0.97	0.97	0.97	423.00
weighted avg	0.97	0.97	0.97	423.00

Source : (Research Results, 2025)

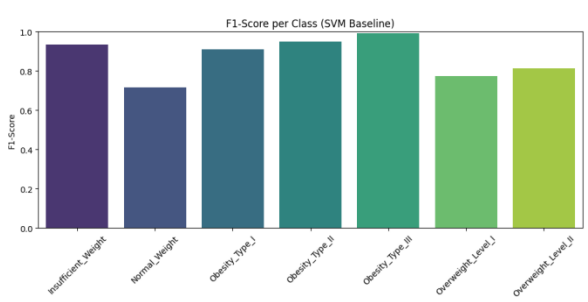
Figure 7. Classification Report SVM Proposed

In contrast, the proposed method, as illustrated in Figure 7, showed a significant improvement across all evaluation metrics. The optimized model reached an accuracy of 97%, with both macro and weighted average F1-scores rising to 0.97. Most notably, all classes achieved F1-scores above 0.93, with several classes such as Obesity_Type_I, Insufficient_Weight, and Obesity_Type_III reaching near-perfect precision and recall values.

These results demonstrate that the application of SMOTE and hyperparameter tuning not only addresses the class imbalance issue but also enhances the model's generalization ability, resulting in a more robust and accurate classification performance across all obesity categories.

F1-Score per Class Visualization

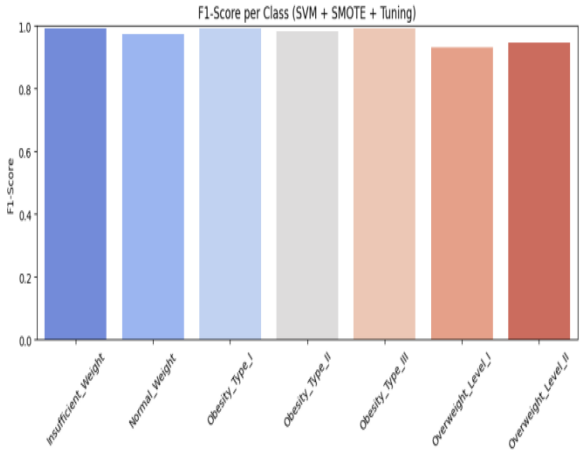
To strengthen the performance analysis of the model in classifying each obesity class, F1-score values per class were visualized for two scenarios: the baseline model and the proposed model (using SMOTE and Hyperparameter Tuning). This visualization helps evaluate how well the model handles data imbalance and accurately classifies each obesity category.



Source : (Research Results, 2025)

Figure 8. F1-Score per Class SVM Baseline

Figure 8 shows the F1-score values of the baseline SVM model without any optimization. It is observed that the model tends to perform better on classes with larger amounts of data, such as *Obesity_Type_III* (0.99) and *Obesity_Type_II* (0.95). On the other hand, relatively low performance is found in *Normal_Weight* (0.72) and *Overweight_Level_I* (0.77), indicating that the baseline model struggles to identify minority classes or more complex patterns.



Source : (Research Results, 2025)

Figure 9. F1-Score per Class SVM Proposed

Figure 9 presents the F1-scores of the SVM model after optimization using the SMOTE method and Hyperparameter Tuning. Overall, there is a significant improvement across all classes. The *Normal_Weight* class, which previously scored only 0.72, increased to 0.97, and *Overweight_Level_I* improved from 0.77 to 0.93. This improvement shows that the proposed method successfully increased the model's sensitivity to minority classes and resulted in more balanced predictions across classes.

Furthermore, the F1-score visualization clearly shows an upward trend across all classes, especially for *Obesity_Type_I* and *Insufficient_Weight*, where the scores improved by more than 10%. This improvement indicates that the decision boundaries learned by the model have become more effective in separating overlapping feature spaces due to optimal values of the kernel parameters. Overall, the visual evidence supports the conclusion that the proposed approach leads to a more balanced and generalizable classifier.

Based on these findings, it can be concluded that applying SMOTE to tackle class imbalance, combined with hyperparameter tuning, significantly improves classification performance, especially in terms of F1-score per class.

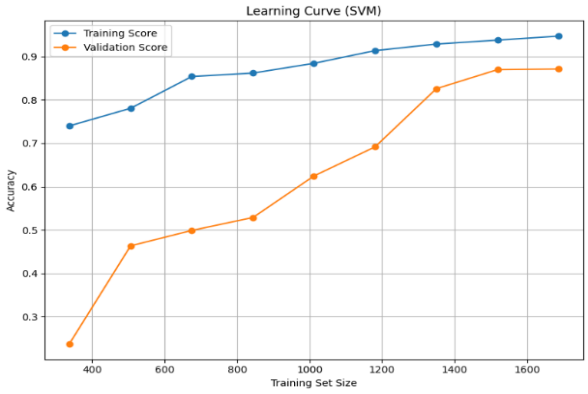
Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i2.6878

The initially lower performance in classes such as Normal_Weight and Overweight_Level_I can be attributed to two main factors. First, these classes had relatively fewer instances compared to others, making them underrepresented during training. Second, these categories share overlapping feature distributions, particularly in variables such as BMI, calorie intake, and physical activity frequency, which can confuse the SVM classifier. After applying SMOTE, the data imbalance was mitigated, allowing the model to receive more training samples for those minority classes. Additionally, hyperparameter tuning via GridSearchCV helped identify optimal values of C and γ, enhancing the decision boundaries and enhances the model's capability to differentiate between closely related classes. These two optimization steps contributed significantly to the improved classification performance observed across all classes.

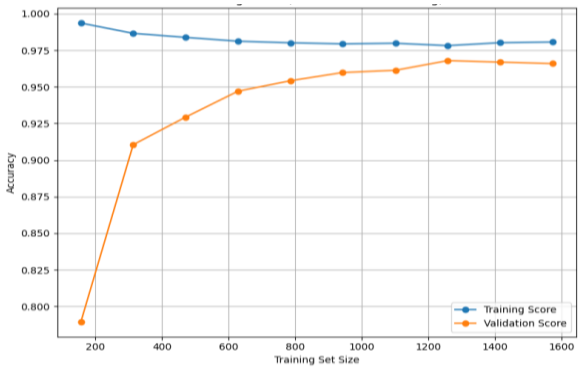
Learning Curve Analysis

In this study, learning curves are utilized to compare the performance of the SVM model before and after the optimization process, which involves balancing the dataset using SMOTE and tuning hyperparameters using GridSearchCV.



Source : (Research Results, 2025)
Figure 10. Learning Curve (SVM)

Figure 10 illustrates the learning curve of the baseline SVM model prior to optimization. The training accuracy shows a steady increase and stabilizes around 0.90, whereas the validation accuracy remains notably lower. Beginning at approximately 0.25, the validation accuracy gradually improves and reaches about 0.86 at the maximum training size. The substantial gap between the training and validation curves signifies overfitting, indicating that the model closely fits the training data but has difficulty generalizing to unseen data.

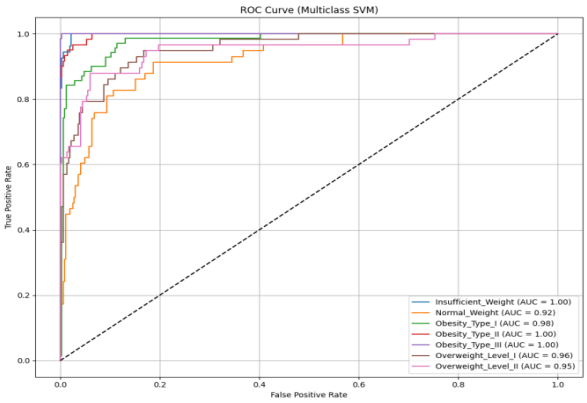


Source : (Research Results, 2025)
Figure 11. Learning Curve (SVM + SMOTE + Tuning)

In contrast, Figure 11 shows the learning curve of the SVM model after the optimization process. The training accuracy remains consistently high, ranging between 0.97 and 0.99. More importantly, the validation accuracy increases rapidly and closely approaches the training curve, reaching over 0.96. The narrow gap between the two curves demonstrates that the optimized model is able to generalize well, significantly reducing overfitting. This indicates that the combination of SMOTE and hyperparameter tuning has effectively enhanced the classification performance of the SVM model for multi-class obesity detection.

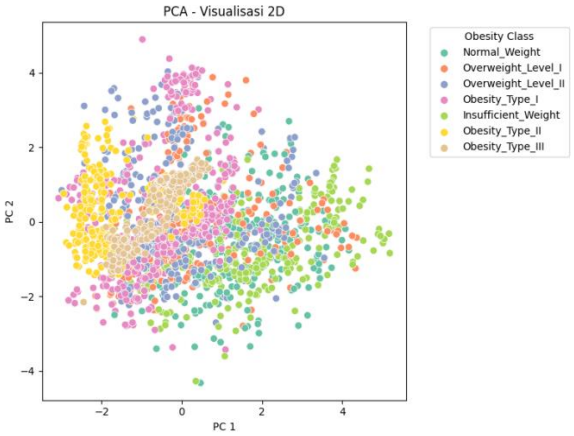
Additional Visualization and Analysis on Baseline Model

In addition to evaluation metrics such as confusion matrix and classification report, this study also presents a series of additional visualizations to further assess and understand the model's performance from different perspectives. These visualizations include the ROC Curve (Multiclass SVM), PCA-based 2D projection, t-SNE-based 2D projection, Learning Curve (SVM), and Validation Curve all of which are generated based on the baseline SVM model.



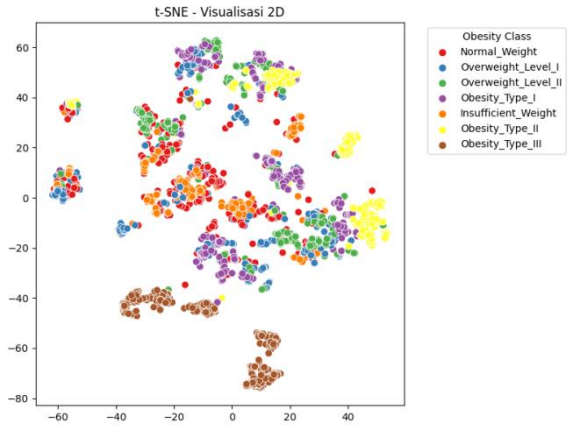
Source : (Research Results, 2025)
Figure 12. ROC Curve (Multiclass SVM)

Figure 12 illustrates the ROC Curve for multiclass classification, displaying the Area Under the Curve (AUC) values for each class. The results demonstrate that the SVM model attained outstanding AUC scores for the majority of classes, with Obesity_Type_II, Obesity_Type_III and Insufficient_Weight achieving an AUC of 1.00. Normal_Weight recorded the lowest AUC of 0.92, indicating relatively lower separability compared to the other classes.



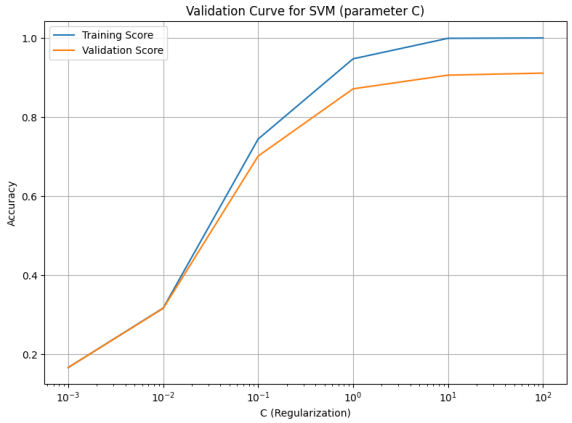
Source : (Research Results, 2025)
Figure 13. PCA-based 2D projection

Figure 13 displays the Principal Component Analysis (PCA) 2D visualization, which projects the original 17-dimensional dataset into two principal components. This visualization helps to observe the clustering tendencies of each obesity class. While some classes show clear grouping, others overlap due to similarities in feature values.



Source : (Research Results, 2025)
Figure 14. t-SNE-based 2D projection

Figure 14 presents the 2D projection using the t-distributed Stochastic Neighbor Embedding (t-SNE) technique. Unlike PCA, t-SNE better captures local structures and non-linear separability among classes. From this plot, the separation between classes appears more distinct, especially for Obesity_Type_III and Obesity_Type_II, which form clearly separated clusters, indicating that the features used were informative for distinguishing those classes.



Source : (Research Results, 2025)
Figure 15. Validation Curve for SVM (parameter C)

Figure 15 displays the validation curve based on the regularization parameter C in SVM. The training and validation scores improve significantly as C increases from 0.001 to 1.0, reaching a near-optimal value at around C = 10. After this point, the training score remains perfect, but the validation score plateaus, indicating potential overfitting if C is too large. This analysis guided the hyperparameter tuning stage in the proposed method.

Model Performance Evaluation

The model's evaluation results, based on precision, accuracy, F1-score, and recall are summarized in Table 5. This table compares the performance of the baseline SVM model with the proposed SVM method, which integrates SMOTE and hyperparameter tuning technique.

The baseline SVM model achieved precision, accuracy, recall, and F1-score of 87%. Conversely, the proposed SVM approach showed significant enhancement, attaining 97% in precision, accuracy, recall, and F1-score. These findings indicate that the proposed method is more effective and robust in classifying obesity levels compared to the baseline model.

Table 7. Model Testing Result

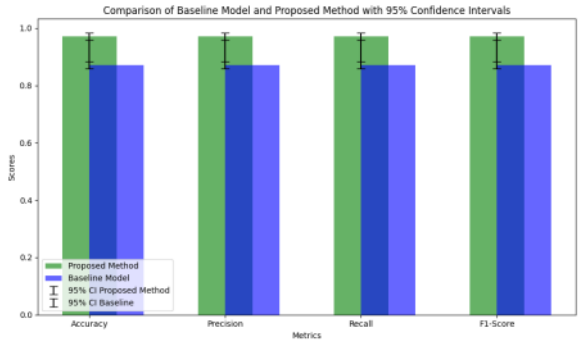
Method	Accuracy	Precision	Recall	F1-Score
SVM Baseline	0,87	0,87	0,87	0,87
SVM proposed method	0,97	0,97	0,97	0,97

Source : (Research Results, 2025)

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i2.6878

As presented in Table 7, the proposed method surpasses the baseline model in all evaluation metrics. The improvement from 87% to 97% in accuracy suggests a notable enhancement of 10%, this outcome highlights the effectiveness of SMOTE in managing class imbalance and the advantages of hyperparameter tuning in optimizing model performance. The uniform improvement across all metrics also demonstrates a well-maintained balance between precision and recall, leading to greater reliability in classification tasks.



Source : (Research Results, 2025)

Figure 16. Confidence Intervals

Figure 16 illustrates the comparison between the baseline and proposed SVM models across four main performance metrics: Accuracy, Precision, Recall, and F1-Score. Each bar represents the score of a model for a specific metric, while the error bars indicate the 95% confidence intervals.

As illustrated in the figure, the proposed approach consistently surpasses the baseline model across all evaluation metrics. Furthermore, the proposed model exhibits narrower confidence intervals, which indicates higher consistency and lower variance in performance across different testing folds. This emphasizes the robustness and generalizability of the proposed approach.

The improvement in performance, along with the reduced uncertainty in predictions (as reflected by the confidence intervals), supports the conclusion that the combination of SMOTE and hyperparameter tuning significantly enhances the classification capability of the SVM model in detecting and categorizing obesity levels.

CONCLUSION

This research sought to enhance the classification performance of obesity levels by optimizing the Support Vector Machine (SVM) algorithm through the application of SMOTE for data balancing and hyperparameter tuning for model optimizationclearly demonstrate that the proposed method significantly outperforms the baseline SVM model in terms of precision, accuracy, recall, and F1-score.

Although the baseline SVM model attained 87% accuracy, 87% precision, 87% recall, and an F1-score of 87%, the proposed method attained better performance across all evaluation metrics, reaching 97% for precision, accuracy, recall, and F1-score. These enhancements demonstrate that the integration of SMOTE and hyperparameter tuning has strengthened the model's generalization capability and improved its effectiveness in addressing class imbalance.

This research confirms that the proposed SVM approach is a more reliable and effective method for multi-class obesity classification. The promising results suggest that this approach can be practically applied in healthcare systems, such as early detection tools for obesity risk levels in routine medical check-ups or as part of digital health monitoring platforms.

To further evaluate the practicality of the model in real-world applications, future work should evaluate its performance on unseen data from diverse populations or different environmental settings. This step is crucial to ensure the model's robustness and generalization across various demographic or geographic contexts. Additionally, future studies may explore integrating wearable sensor data, dietary or activity logs, and evaluating the model's deployment in clinical or remote healthcare settings to enhance its practicality and robustness. Furthermore, feature importance analysis using model-agnostic interpretability methods, such as SHAP or permutation importance, will be incorporated to better understand the contribution of individual features to the model's predictive performance.

ACKNOWLEDGEMENT

We sincerely extend our gratitude to the Ministry of Higher Education, Science, and Technology (Kemdiktisaintek) of the Republic of Indonesia, Directorate General of Research and Development, for funding this research under the Master's Thesis Postgraduate Research (PPTM) scheme for the 2025 fiscal year. With the Master Contract Number: 122/C3/DT.05.00/PL/2025, dated June 11, 2025. And the Derivative Contract Number: 77/SPK/LL1/AL.04.03/PL/2025.

REFERENCE

[1] D. Mohajan and H. K. Mohajan, "Obesity and Its Related Diseases: A New Escalating

Alarming in Global Health," *J. Innov. Med. Res.*, vol. 2, no. 3, pp. 12–23, 2023, doi: 10.56397/jimr/2023.03.04.

[2] A. Koceva, R. Herman, A. Janez, M. Rakusa, and M. Jensterle, "Sex- and Gender-Related Differences in Obesity: From Pathophysiological Mechanisms to Clinical Implications," *Int. J. Mol. Sci.*, vol. 25, no. 13, 2024, doi: 10.3390/ijms25137342.

[3] A. Gutiérrez-Gallego *et al.*, "Combination of Machine Learning Techniques to Predict Overweight/Obesity in Adults," *J. Pers. Med.*, vol. 14, no. 8, 2024, doi: 10.3390/jpm14080816.

[4] Nikolaos Tzenios, "Obesity As a Risk Factor for Different Types of Cancer," *EPRA Int. J. Multidiscip. Res.*, no. February, pp. 97–100, 2023, doi: 10.36713/epra12421.

[5] Y. Zhou, S. Wang, Y. Xie, T. Zhu, and C. Fernández, "An Improved Particle Swarm Optimization-Least Squares Support Vector Machine-Unscented Kalman Filtering Algorithm on SOC Estimation of Lithium-Ion Battery," *Int. J. Green Energy*, vol. 21, no. 2, pp. 376–386, 2023, doi: 10.1080/15435075.2023.2196328.

[6] S. Jeon, M. Kim, J. Yoon, S. Lee, and S. Youm, "Machine learning-based obesity classification considering 3D body scanner measurements," *Sci. Rep.*, vol. 13, no. 1, pp. 1–10, 2023, doi: 10.1038/s41598-023-30434-0.

[7] A. Maulana, R. Perucha, F. Afidh, N. B. Maulydia, and G. M. Idroes, "Predicting Obesity Levels with High Accuracy : Insights from a CatBoost Machine Learning Model," vol. 2, no. 1, pp. 17–27, 2024, doi: 10.60084/ijds.v2i1.195.

[8] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets : an investigation of performance for machine learning and feature engineering," *J. Big Data*, 2024, doi: 10.1186/s40537-024-00943-4.

[9] A. I. Putri, N. A. Husna, N. M. Cia, and M. A. Arba, "Implementation of K-Nearest Neighbors , Naïve Bayes Classifier , Support Vector Machine and Decision Tree Algorithms for Obesity Risk Prediction," *Inst. Res. Publ. Indones. Public Res. J. Eng. Data Technol. Comput. Sci.*, vol. 2, no. July, pp. 26–33, 2024.

[10] N. Khodadadi, M. Saber, and M. Abotaleb, "A Data-Driven Approach for Obesity Classification using Machine Learning," *J. Artif. Intell. Metaheuristics*, vol. 3, no. 2, pp. 08–17, 2023, doi: 10.54216/jaim.030201.

[11] D. A. Kristiyanti, S. A. Sanjaya, V. C. Tjokro, and J. Suhali, "Dealing imbalance dataset problem in sentiment analysis of recession in Indonesia," *IAES Int. J. Artif. Intell.*, vol. 13, no. 2, pp. 2058–2070, 2024, doi: 10.11591/ijai.v13.i2.pp2060-2072.

[12] S. A. Alsareii *et al.*, "Machine-Learning-Enabled Obesity Level Prediction Through Electronic Health Records," *Comput. Syst. Sci. Eng.*, 2023, doi: 10.32604/csse.2023.035687.

[13] R. Ruliana, Z. Rais, M. Marni, and A. S. Ahmar, "Implementation of the Support Vector Regression (SVR) Method in Inflation Prediction in Makassar City," *ARRUS J. Math. Appl. Sci.*, vol. 4, no. 1, pp. 28–35, 2024, doi: 10.35877/mathscience2608.

[14] M. Y. Anshori, T. Herlambang, and M. F. A. Yaziz, "Optimalizing Occupancy of Hospitality Sector Using Support Vector Regression and Genetic Algorithm," *J. Revenue Pricing Manag.*, no. 0123456789, 2025, doi: 10.1057/s41272-025-00539-4.

[15] O. Iparraguirre-Villanueva, L. Mirano-Portilla, M. Gamarra-Mendoza, and W. Robles-Espiritu, "Predicting Obesity in Nutritional Patients using Decision Tree Modeling," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 3, pp. 254–260, 2024, doi: 10.14569/IJACSA.2024.0150326.

[16] Z. Azam, M. Islam, and M. N. Huda, "Comparative Analysis of Intrusion Detection Systems and Machine Learning-Based Model Analysis Through Decision Tree," *IEEE Access*, vol. 11, no. June, 2023.

[17] R. Hassanzadeh, M. Farhadian, and H. Rafieemehr, "Hospital mortality prediction in traumatic injuries patients : comparing different SMOTE - based machine learning algorithms," *BMC Med. Res. Methodol.*, pp. 1–15, 2023, doi: 10.1186/s12874-023-01920-w.

[18] X. Chuhan, Pablo, and X. Jiang, "Empirical Study of Overfitting in Deep Learning for Predicting," *Cancers (Basel).*, 2023.

[19] W. Gou, H. Zhang, and R. Zhang, "Multi-Classification and Tree-Based Ensemble Network for the," 2023.

[20] S. N. Khan, S. U. Khan, H. Aznaoui, and C. B. Şahin, "Generalization of linear and non-linear support vector machine in multiple fields : a review," *Comput. Sci. Inf. Technol.*, vol. 4, no. 3, pp. 226–239, 2023, doi: 10.11591/csit.v4i3.pp226-239.

[21] A. Sultana and R. Islam, "Machine learning framework with feature selection

VOL. 11. NO. 2 NOVEMBER 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i2.6878

approaches for thyroid disease classification and associated risk factors identification," *J. Electr. Syst. Inf. Technol.*, 2023, doi: 10.1186/s43067-023-00101-5.

[22] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, *Handling imbalanced medical datasets : review of a decade of research*, vol. 57, no. 10. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10884-2.

[23] J. Rusman, B. Z. Haryati, and A. Michael, "Optimisasi Hiperparameter Tuning pada Metode Support Vector Machine untuk Klasifikasi Tingkat Kematangan Buah Kopi," *J. Komput. dan Inform.*, vol. 11, no. 2, pp. 195–202, 2023, doi: 10.35508/jicon.v11i2.12571.

[24] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Inf.*, vol. 14, no. 1, 2023, doi: 10.3390/info14010054.

[25] T. Elansari, M. Ouanan, and H. Bourray, "Mixed Radial Basis Function Neural Network Training Using Genetic Algorithm," *Neural Process. Lett.*, vol. 55, no. 8, pp. 10569–10587, 2023, doi: 10.1007/s11063-023-11339-5.

[26] D. M. Belete and M. D. Huchaiah, "Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results," *Int. J. Comput. Appl.*, vol. 44, no. 9, pp. 875–886, 2022, doi: 10.1080/1206212X.2021.1974663.

[27] C. Gyurik, D. Van Vreumingen, and V. Dunjko, "Structural risk minimization for quantum linear classifiers," 2022.

[28] O. Farombi, "DEVELOPMENT OF KERNEL SVM MODEL TO PREDICT CUSTOMERS SUV DEVELOPMENT OF KERNEL SVM MODEL TO PREDICT," no. July, 2024, doi: 10.13140/RG.2.2.13774.47689.

[29] G. Boteju, L. Tang, and M. S. Brown, "Support Vector Machine for Predicting Student Dropout Under Different Normalization Methods," *Proc. - 2024 IEEE Int. Conf. Big Data, BigData 2024*, no. December, pp. 8633–8636, 2024, doi: 10.1109/BigData62323.2024.10825023.

[30] S. Panno *et al.*, "A review of the most common and economically important diseases that undermine the cultivation of tomato crop in the mediterranean basin," *Agronomy*, vol. 11, no. 11, pp. 1–45, 2021, doi: 10.3390/agronomy11112188.

[31] B. Gaye, D. Zhang, and A. Wulamu, "Improvement of Support Vector Machine Algorithm in Big Data Background," *Math. Probl. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/5594899.

[32] Jumanto *et al.*, "Optimizing Support Vector Machine Performance for Parkinson's Disease Diagnosis Using GridSearchCV and PCA-Based Feature Extraction," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 10, no. 1, pp. 38–50, 2024, doi: 10.20473/jisebi.10.1.38-50.

[33] Smita Panigrahy, "SMOTE-based Deep LSTM System with GridSearchCV Optimization for Intelligent Diabetes Diagnosis," *J. Electr. Syst.*, vol. 20, no. 7s, pp. 804–815, 2024, doi: 10.52783/jes.3455.