IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.19, No.2, April 2025, pp. 141~152
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: 10.22146/ijccs.103994

SMOTE-SVM for Handling Imbalanced Data in Obesity Classification

Farhan Radhiansyah Razak[1], Muhammad Kunta Biddinika*[2], Herman Yuliansyah[3] , Dewi Soyusiawaty[4]
[1,2]Master Program of Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[3,4]Department of Informatics, Universitas Ahmad Dahlan,Yogyakarta, Indonesia
e-mail: [1]farhanrrazak@gmail.com, ***[2]muhammad.kunta@mti.uad.ac.id**,
[3]herman.yuliasnyah@tif.uad.ac.id, [4]dewi.soyusiawaty@tif.uad.ac.id

Abstrak

Obesitas merupakan masalah kesehatan yang signifikan yang terkait dengan berbagai penyakit kronis, sehingga klasifikasi dini menjadi sangat penting untuk intervensi yang efektif. Studi ini bertujuan untuk meningkatkan akurasi klasifikasi obesitas pada dataset yang tidak seimbang dengan mengintegrasikan teknik Synthetic Minority Over-sampling Technique (SMOTE) dengan algoritma Support Vector Machine (SVM) menggunakan kernel Radial Basis Function (RBF) dan Linear. Ketidakseimbangan data diatasi dengan menerapkan SMOTE dan Random Undersampling (RUS) untuk mengevaluasi performa model dalam berbagai kondisi. Temuan penelitian menunjukkan bahwa teknik penyeimbangan secara substansial meningkatkan kinerja klasifikasi, dengan model SVM Linear mencapai akurasi tertinggi sebesar 96,54% saat data diseimbangkan menggunakan SMOTE. Selain itu, hasil juga menunjukkan bahwa SMOTE lebih unggul dibandingkan RUS dalam mempertahankan informasi penting, dan bahwa model kernel Linear lebih robust terhadap dataset ini dibandingkan dengan kernel RBF. Temuan ini menekankan pentingnya penanganan ketidakseimbangan data dalam tugas klasifikasi yang berkaitan dengan kesehatan untuk memastikan prediksi yang adil dan akurat.

Kata kunci*—Obesitas, SMOTE, RUS, RBF, Linear*

Abstract

Obesity is a significant health issue associated with various chronic diseases, making early classification essential for effective interventions. This study aims to enhance obesity classification accuracy on imbalanced datasets by integrating the Synthetic Minority Over-sampling Technique (SMOTE) with Support Vector Machine (SVM) classifiers using Radial Basis Function (RBF) and Linear kernels. The dataset imbalance was addressed using SMOTE and Random Undersampling (RUS) to evaluate the models under different conditions. The findings demonstrate that balancing techniques substantially improve classification performance, with the Linear SVM model achieving the highest accuracy of 96.54% when balanced using SMOTE. Moreover, results indicate that SMOTE outperforms RUS by preserving more information, and that the Linear kernel model is more robust for this dataset compared to the RBF kernel. These insights highlight the importance of handling data imbalance in health-related classification tasks to ensure fairness and accuracy.

Keywords—*Obesity, SMOTE, RUS, RBF, Linear*

Received January 19th,2025; Revised April 28th, 2025; Accepted April 30th, 2025

1. INTRODUCTION

Obesity is a medical condition defined as disproportionate fat storage in the body that might adversely affect health [1]. Over time, a major cause of health problems is eating more calories than the body uses. The extra calories are stored as fat, leading to various health issues[2]. A sedentary lifestyle, fast-paced work life, and changing eating habits are causing many adults to have difficulty controlling their weight [3], [4]. In recent years, obesity become a major health issue affecting people all over the world [5], [6]. This is why it has become a serious issue that harms a person's physical health, lowers their quality of life, and shortens their overall life expectancy [7]. Obesity negatively impacts many parts of the body, including the endocrine system, heart and blood vessels, lungs, digestive system, skin, urinary and reproductive systems, and bones and muscles [8]. The World Health Organization (WHO) states that obesity leads to serious health problems, including heart diseases, diabetes, cancer, and various musculoskeletal disorders [9], [10]. According to WHO, over 2.8 million adults die every year from health issues caused by being overweight or obese [11]. The Public Health Agency of Turkiye has found that being overweight is a significant health concern, causing over 1 million deaths annually across the European Region [12].

Machine Learning algorithms have transformed the healthcare industry by providing advanced classification methods. These techniques are widely used for tasks like diagnosing diseases, predicting health risks, and recommending effective treatments. These algorithms leverage large datasets to identify patterns and relationships within medical data, allowing for more accurate and efficient decision-making. They have demonstrated remarkable power in distinguishing between disease states, stratifying patients based on risk profiles, and optimizing treatment strategies [13], [14]

Various studies have explored the use of machine learning algorithms in predicting obesity risk with various approaches. Study [15] comparing KNN, Naïve Bayes, and SVM algorithms found that Decision Trees achieved the highest accuracy at 84.98% for specific datasets. Study [16] focused on predicting obesity in adults, utilizing algorithms such as Logistic Regression, Classification and Regression Trees, and Naïve Bayes. Logistic Regression emerged as the best performer with an accuracy of 72% and an AUC of 79%, supported by SMOTE to address data imbalance. Study [17] investigated the effect of physical activity on obesity prediction. They tested a wide array of algorithms, including Naïve Bayes, RBF, KNN, CVR, and others. The Random Subspace algorithm provided the best result with an accuracy of 67% and an AUC of 64%. Study [18] aimed to identify risk factors predicting obesity using machine learning classifiers. Multi-Layer Perceptron (MLP) was the top performer with an AUC of 78%, among other tested models like Random Forest, SVM, and Logistic Regression. Meanwhile, Study [19] applied various ML techniques to build a model for identifying a person's obesity. Decision Tree achieved the highest accuracy of 78%, outperforming other models like SVM, KNN, and Gradient Boosting.

Although various algorithms have been used, obesity prediction still faces two main challenges that need to be addressed: the complexity of the relationships between variables influencing obesity, which are often non-linear, and the data imbalance between healthy individuals and those with obesity, where the number of healthy individuals is much larger, causing machine learning models to be biased towards the majority class. To address this issue, various data balancing techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and Random Undersampling (RUS) are often applied [20]. SMOTE works by generating synthetic data for the minority class, while RUS reduces the amount of data from the majority class, aiming to create a more balanced data distribution and prevent model bias. The use of these two techniques can improve model accuracy and assist in the development of more effective intervention strategies.

This study aims to explore the use of SVM with linear and Radial Basis Function (RBF) kernels in predicting obesity, focusing on the development of a more accurate model through the application of data balancing techniques. The combination of linear and RBF kernels allows for

handling data with complex non-linear relationships, while the implementation of SMOTE and RUS is expected to address the class imbalance issue in obesity data. With this approach, it is hoped that the resulting model can provide more reliable results in predicting obesity and its risk factors, as well as support the development of more targeted interventions.

The main contribution of this research is the application of machine learning techniques using SVM, which utilizes both types of kernels to handle complex data. In addition, this study uses data balancing methods to enhance the accuracy of obesity classification, which plays a crucial role in creating better and more reliable models for obesity classification. Therefore, this study specifically aims to develop a robust and accurate obesity classification model by applying SMOTE to handle data imbalance and utilizing Support Vector Machine (SVM) classifiers with Radial Basis Function (RBF) and Linear kernels. The goal is to improve predictive performance and ensure balanced representation across all obesity classes, addressing critical challenges often encountered in healthcare datasets.

2. METHODS

The aim of this research is to explore the issues related to data imbalance in obesity classification datasets.

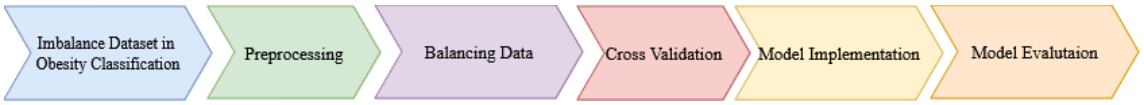

Figure 1. Research Implementation

As shown in Figure 1, this study involves several steps. First, preprocessing is done on the imbalanced obesity data from companies. During this step, data is labeled and scaled. Next, cross-validation is applied using 10-fold validation. Afterward, the classification process is carried out using SVM algorithms with both Radial Basis Function and Linear Kernels. Finally, the performance of the classification model is evaluated to determine its effectiveness in predicting obesity.

2.1 Imbalance Dataset in Obesity Classification

The dataset used in this study is presented in Table 1. It contains the features and variables essential for the classification of obesity, providing the foundation for the analysis and model development.

Table 1. Obesity Dataset

Gender	Age	Height	Weight	Family History Overweight	FAFC	FCVC	NCP	CAEC	SMOKE	CH20	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
Female	21	1.62	64.00	Yes	No	2.0	3.0	Sometimes	No	2.00	No	0,00	1.00	No	Public Transport	Normal Weight
Female	21	1.52	56.00	Yes	No	3.0	3.0	Sometimes	Yes	2.00	Yes	3.00	0.00	Sometimes	Public Transport	Normal Weight
Male	23	1.80	77.00	Yes	No	2.0	3.0	Sometimes	No	2.00	No	2.00	1.00	Frequently	Public Transpor	Normal Weight
Male	27	1.80	87.00	No	No	3.0	3.0	Sometimes	No	2.00	No	2.00	0.00	Frequently	Walking	Overweight Level I
Male	22	1.80	89.00	No	No	2.0	1.0	Sometimes	No	2.00	No	0.00	0.00	Sometimes	Public Transport	Overweight Level 2
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
Female	20	1.71	131.40	Yes	Yes	3.0	3.0	Sometimes	No	1.72	No	1.67	0.90	Sometimes	Public Transport	Obesity Type 3
Female	21	1.74	133.74	Yes	Yes	3.0	3.0	Sometimes	No	2.00	No	1.34	0.59	Sometimes	Public Transport	Obesity Type 3
Female	22	1.75	133.68	Yes	Yes	3.0	3.0	Sometimes	No	2.05	No	1.41	0.64	Sometimes	Public Transport	Obesity Type 3
Female	24	1.73	133.34	Yes	Yes	3.0	3.0	Sometimes	No	2.85	No	1.13	0.58	Sometimes	Public Transport	Obesity Type 3
Female	23	1.73	133.47	Yes	Yes	3.0	3.0	Sometimes	No	2.86	No	1.02	0.71	Sometimes	Public Transport	Obesity Type 3

Table 1 presents the initial dataset used for the classification of obesity. Data imbalance is one of the primary challenges in obesity classification. This issue arises when the number of samples in the majority class significantly outweighs those in the minority class, such as individuals with insufficient weight (272 samples) or higher obesity levels like Obesity Type 3 (324 samples). The dataset used in this study, which originates from Kaggle, consists of seven weight-based categories: Insufficient Weight (272 samples), Normal Weight (287 samples), Overweight Level 1 (290 samples), Overweight Level 2 (290 samples), Obesity Type 1 (351 samples), Obesity Type 2 (297 samples), and Obesity Type 3 (324 samples), totaling 2,111 samples. This imbalance can cause machine learning models to favor the majority class, leading to reduced accuracy for the minority classes, which are often critical for health-related interventions. Data balancing techniques such as SMOTE and RUS are employed to address this. SMOTE generates synthetic samples for minority classes by interpolating existing data points, while RUS reduces the size of the majority class by selecting a representative subset.

These techniques help mitigate bias, improve classification performance for minority classes, and ensure fairer and more accurate predictions, thereby supporting the development of effective intervention strategies.

2. 2 Preprocessing

During this stage, several steps are taken to ensure the data is properly prepared for modeling. These steps include removing any empty or duplicate entries, scaling the data into numerical values using a standard scaler, and labeling the data correctly. Labeling is crucial because it defines and organizes each data point, allowing the data to be grouped based on similar features. This is particularly important for processes like SMOTE and RUS, which are used for balancing the data and will be explained in the next section. For example, the preprocessing steps applied to the data in Table 1 are shown in Table 2.

Table 2. Preprocessing Results

Gender	Age	Height	Weight	Family History Overweight	FAFC	FCVC	NCP	CAEC	SMOKE	CH20	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
0	21	1.62	64.00	1	0	2.0	3.0	2	0	2.00	0	0,00	1.00	3	3	1
0	21	1.52	56.00	1	0	3.0	3.0	2	1	2.00	1	3.00	0.00	2	3	1
1	23	1.80	77.00	0	0	2.0	3.0	2	0	2.00	0	2.00	1.00	1	3	1
1	27	1.80	87.00	0	0	3.0	3.0	2	0	2.00	0	2.00	0.00	1	4	5
1	22	1.80	89.00	0	0	2.0	1.0	2	0	2.00	0	0.00	0.00	2	3	6
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
0	20	1.71	131.40	1	1	3.0	3.0	2	0	1.72	0	1.67	0.90	2	3	4
0	21	1.74	133.74	1	1	3.0	3.0	2	0	2.00	0	1.34	0.59	2	3	4
0	22	1.75	133.68	1	1	3.0	3.0	2	0	2.05	0	1.41	0.64	2	3	4
0	24	1.73	133.34	1	1	3.0	3.0	2	0	2.85	0	1.13	0.58	2	3	4
0	23	1.73	133.47	1	1	3.0	3.0	2	0	2.86	0	1.02	0.71	2	3	4

2. 3 Balancing Data

Data balancing is a technique used to address class imbalances in a dataset to ensure that machine learning models are not biased toward the majority class. The following are the techniques commonly used to handle data imbalance:

a. Using Syntetic Minority Oversampling (SMOTE)

SMOTE is an effective technique for handling class imbalance in datasets by creating synthetic examples of the underrepresented class [21]. It works by generating new data points for the minority class through interpolation between existing instances of that class, helping to create a more balanced dataset. SMOTE helps to mitigate the problem of biased classification models that tend to favor the majority class due to its higher representation in the dataset. By introducing syntethic samples, SMOTE enhances the diversity of the minority class, allowing machine learning algoritms to better learn the underlying patterns and improve

classification performance. This technique has been widely adopted in various fields, including healthcare [22], [23], finance [24], [25], and image recognition [26], [27], this method is particularly useful in situations where datasets are imbalanced, leading to more accurate and reliable predictive models.

b. Using Random Undersampling (RUS)

Undersampling is one of the simplest methods for addressing unbalanced data. Random undersampling involves calculating the difference in the number of instances between the majority and minority classes. Next, random instances are selected and removed from the majority class until the number of instances in the majority class equals the number in the minority class. This technique aims to balance the distribution of data across classes. Random undersampling helps prevent machine learning models from being biased toward the majority class by reducing the number of instances in that class. This process improves the model's ability to accurately classify instances from the minority class [28].

2. 4 Cross Validation

Cross-validation is a crucial step to ensure that the model's performance is robust and not biased by the specific partition of training and testing data. In this study, a 10-fold cross-validation technique was used, where the dataset was divided into 10 equal subsets. For each iteration, nine subsets were used for training the model, and the remaining one subset was used for testing. This process was repeated 10 times as shown in Figure 2, ensuring that each subset was used for testing exactly once. By averaging the results across all iterations, cross-validation provides a reliable estimate of the model's generalization capability [29]. This approach mitigates the risk of overfitting, particularly when working with imbalanced datasets, and ensures that the model's performance metrics, such as accuracy, precision, recall, and F1-score, reflect its true predictive power across diverse data distributions.

Figure 2. Cross Validation

2. 5 Model Implementation

Support Vector Machine (SVM) is one of the machine learning methods used for classification. SVM converts data into a higher-dimensional space to find a hyperplane that

separates different classes with the largest margin. In the context of SVM with kernels, there are two types of kernels that are often used:

a. Radial Basis Function

The Radial Basis Function (RBF) kernel is useful when the data is unevenly distributed. When using RBF for training, two key parameters need to be considered, C and gamma. The C parameter controls how much the model tries to avoid errors when classifying the training data. A higher C value reduces the misclassification of the training data. The gamma parameter determines the range of influence a single training data point has. A smaller gamma value means that the influence of each data point reaches farther, affecting a larger area. Equation (1) is used for RBF kernel [30].

$$K(x, y) = \exp(-\gamma \,||\, x - y \,||^2) \quad (1)$$

b. Linear

The linear kernel, also known as a soft margin, aims to find a straight line (hyperplane) that best separates the data. However, it allows for some misclassifications. Even though it permits a few errors, the linear kernel still strives to find a line that maximizes the margin and minimizes misclassification. The level of tolerance for misclassification significantly impacts the accuracy of the hyperplane. In sklearn, this tolerance is controlled by the C parameter. A higher C value means less tolerance for misclassifications and results in a narrower margin. Equation (2) is used for linear kernel [30].

$$K(x,y) = x.y + C \quad (2)$$

2. 6 Model Evaluation

The outcomes of the analytical models are examined in this part according to assessment metrics, including accuracy, precision, recall, and F-measure. Furthermore, a rationale is given for the choice of model that is best for predicting obesity levels based on the comparison that was made. The effectiveness of SVM, NB, and KNN is evaluated using four categories. They are False Positive (FP), True Negative (TN), True Positive (TP), and False Negative (FN) [31].

Accuracy shows how often the model gets things right. It's the total correct predictions divided by all predictions. Equation (3) is used for accuracy [31].

$$Accuracy = \frac{TP + T\,N}{TP + TN + FP + FN} \quad (3)$$

Precision is when the model predicts positively and how often the prediction is correct. Equation (4) is used in precision [31].

$$Precision = \frac{TP}{FP + TP} \quad (4)$$

Recall is when the actual class is positive and how much the model predicts it to bes positive. Equation (5) is used in the recall [31].

$$Recall = \frac{TP}{FN + TP} \quad (5)$$

Equation (6) combines precision and recall into one score to give a balanced view. It's useful when we care about both finding positives and being accurate [31].

$$F1 - Score = \frac{2 \text{ x precision x recall}}{\text{precision + recall}} \quad (6)$$

3. RESULTS AND DISCUSSION

In this study, different scenarios were tested to explore the effects of oversampling with SMOTE and undersampling with RUS on class imbalance. The goal was to understand how these techniques impact the balance of datasets. The study used both RBF and Linear SVM models as classifiers to analyze imbalanced and balanced datasets after applying SMOTE and RUS. For each scenario, the data was evaluated using 10-fold cross-validation, as explained earlier, to ensure reliable model performance and reduce the risk of overfitting.

3.1 Handling Data of Data Imbalance

At this stage, the researcher applies SMOTE and RUS to address the issue of imbalanced data. SMOTE increases the number of instances in the minority class to balance the dataset with the majority class, while RUS reduces the number of instances in the majority class to correct the imbalance. As shown in Figure 6, the Obesity Classification dataset had an imbalance, with the following number of instances in each class: Insufficient Weight (272), Normal Weight (287), Overweight Level 1 (290), Overweight Level 2 (290), Obesity Type 1 (351), Obesity Type 2 (297), and Obesity Type 3 (324).

As shown in Figure 3 and Figure 4, the number of instances in the minority class has increased after applying SMOTE and RUS. This shows that both techniques have successfully created synthetic samples, balancing the dataset with the majority class. As a result, the Obesity dataset is ready for use in developing machine learning models.

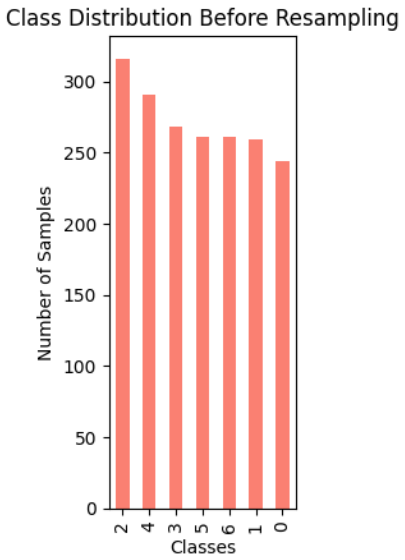

Figure 3. Imbalance Data

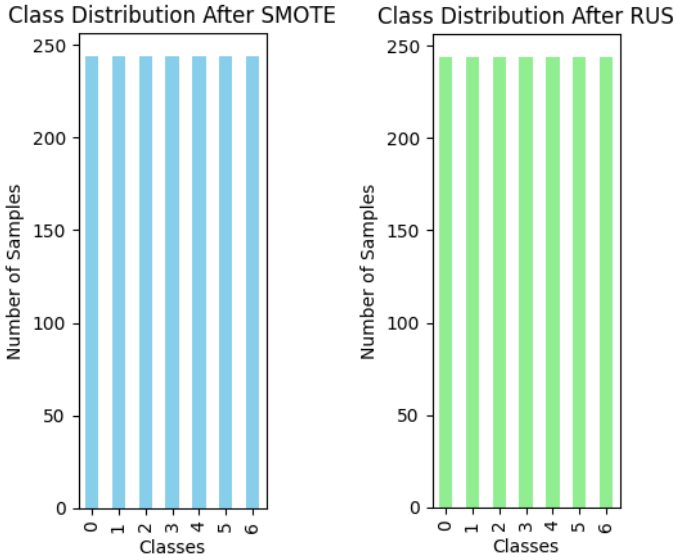

Figure 4. Balance Data with SMOTE and RUS

3.2 Testing Scenarios

The testing scenarios are shown in Table 3. This testing involved using both the Radial Basis Function (RBF) kernel and the Linear Model of SVM for the modeling process. The tests were carried out using both imbalanced and balanced datasets.

Table 3. Testing Scenarios

No	Testing Data	Kernel
1	Imbalanced Data	Radial Basis Function, Linear
2	Balanced Data with SMOTE	Radial Basis Function, Linear
3	Balanced Data with RUS	Radial Basis Function, Linear

3.3 Testing Result

Integrating SMOTE for oversampling and RUS for undersampling has a positive impact on the classification performance of the obesity dataset. As shown in Table 4, the RBF model with imbalanced data achieves an accuracy of 89%, while the Linear model reaches 96%. These results suggest that both models work well with imbalanced data but perform even better when data balancing techniques are applied.

Table 4. Model Performances

Kernel	Data Balancing Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RBF	Imbalanced Data	89	89	89	89
	Balanced with SMOTE	93.84	94.10	93.84	93.84
	Balanced with RUS	93.13	93.49	93.13	93.15
Linear	Imbalanced Data	96	96	96	96
	Balanced with SMOTE	96.54	96.66	96.54	96.54
	Balanced with RUS	96.44	96.58	96.44	96.44

After applying SMOTE, the RBF model's accuracy improves by 4.84%, with corresponding increases in precision, recall, and F1-score of 4.10%, 4.84%, and 4.84%, respectively. Similarly, the Linear model shows an accuracy improvement of 0.54%, with slight enhancements across all metrics. When using RUS, the RBF model's accuracy increases by 4.13%, while the Linear model improves by 0.44%, maintaining consistent performance across evaluation metrics. These results highlight the effectiveness of SMOTE and RUS in enhancing classification accuracy and reducing bias toward the majority class.

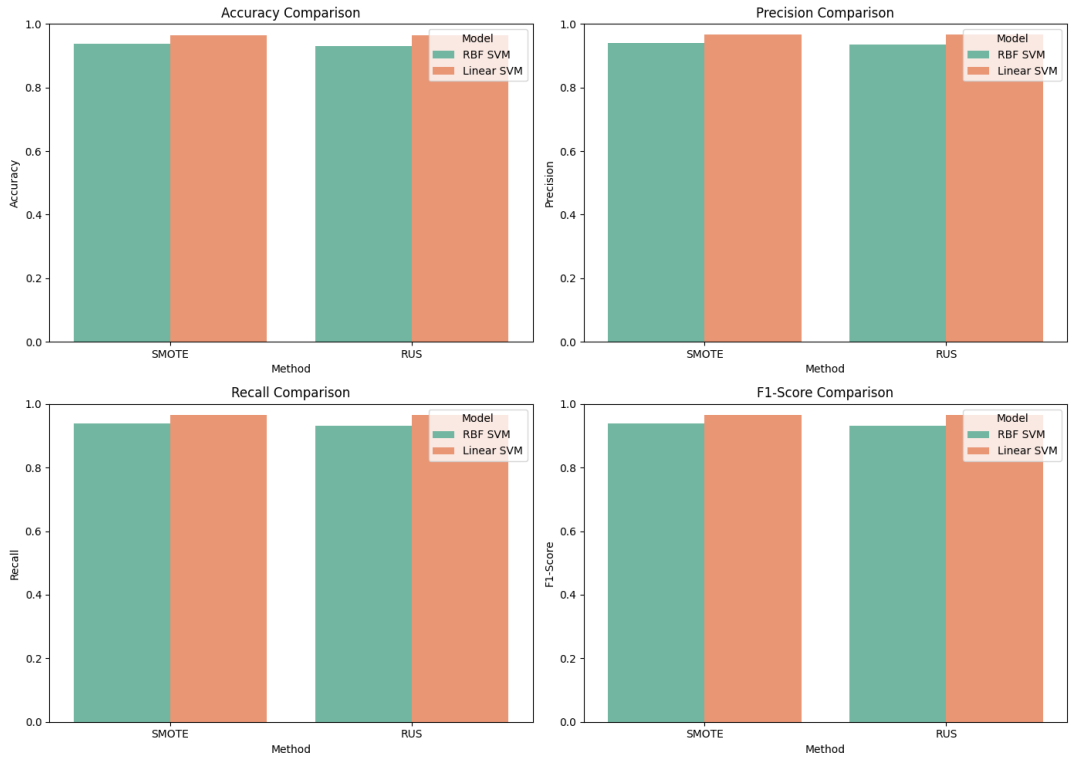

Figure 5. Model Performance SMOTE and RUS

The performance of each scenario is shown in Figure 5. Both the Radial Basis Function and Linear models produce more satisfactory results when processed with SMOTE and RUS. This can be seen in Figure 3, where the Obesity Classification data is unevenly distributed. This result supports the findings from previous studies [32], which suggest that if the class imbalance is not addressed, the model's performance will suffer. Both the Radial Basis Function and Linear models showed improved performance when applied to datasets balanced using SMOTE and RUS, resolving the issue of imbalanced data. This aligns with research conclusions [32], highlighting that fixing class imbalance significantly boosts performance. The use of SMOTE and RUS to handle imbalanced data proves to be an effective method for improving model performance on obesity datasets. This demonstrates that the data balancing process successfully enhanced model accuracy. Addressing class imbalance is crucial as it ensures that all categories are properly represented during training, resulting in more accurate and fair predictions. The findings of this study show an improvement in the performance of each model. However, The Linear model balanced with SMOTE performs better than RUS and RBF kernel methods because SMOTE creates high-quality synthetic samples for the minority class, helping the Linear model better understand the relationships in the data. Unlike the RBF kernel, which can be sensitive to noise and requires more tuning, the Linear model is simpler, faster, and works well with linearly separable data. This combination makes SMOTE with the Linear model a powerful and reliable choice for addressing class imbalance, ensuring better representation and fairer predictions across all classes.

The superior performance of the Linear SVM model combined with SMOTE can be explained by the characteristics of the dataset used in this study. After balancing, the data distribution becomes more linear, enabling the Linear kernel to distinguish between classes more effectively without experiencing overfitting. In contrast, the RBF kernel, which maps data into higher dimensions, becomes more sensitive to noise introduced during the oversampling process, resulting in slightly lower generalization capability. Furthermore, SMOTE has proven to be a more effective balancing technique compared to RUS, as it enriches the minority class without sacrificing valuable samples from the majority class, thereby preserving critical information necessary for accurate classification. Although SMOTE significantly improved the overall performance of the model, there remains a potential risk of overfitting due to synthetic data. Therefore, future research may consider applying techniques such as Grid Search Cross Validation (Grid CV) to optimize the model's hyperparameters, aiming to achieve more accurate classification results and better generalization. These findings emphasize that proper data balancing and model selection are crucial when handling imbalanced datasets, particularly in sensitive domains such as healthcare.

3.4 Comparison to Other Studies

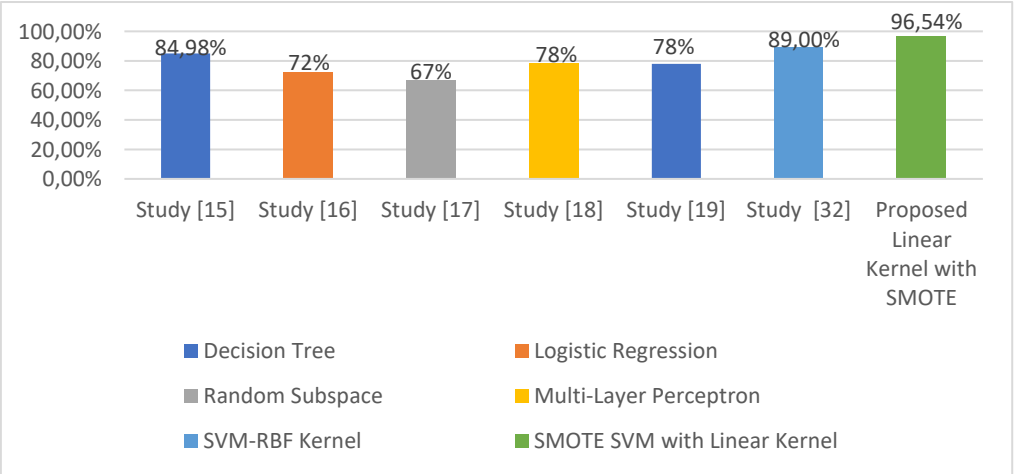

Figure 6. Comparison of the Proposed Model with Other Studies

Based on Figure 6, the proposed model utilizing SVM linear kernel and SMOTE, achieves an accuracy of 96.54%, making it the most effective among the compared studies. This accuracy surpasses the results of study [15] achieved 84.98% using a Decision Tree, and study [18], [19] both reported 78% accuracy using Multi-Layer Perceptron and Decision Tree models, respectively. Additionally, study [16] reached 72% accuracy with Logistic Regression, and study [17] achieved only 67% using the Random Subspace method. The significant leap in accuracy indicates that the proposed model delivers a substantial improvement over previous studies. This improvement likely stems from the integration of SVM, which excels in handling complex decision boundaries, and SMOTE, which effectively addresses class imbalance issues, enhancing the model's ability to generalize and deliver precise predictions.

4. CONCLUSIONS

The results of this study demonstrate that applying data balancing techniques such as SMOTE and RUS effectively improves the performance of machine learning models in obesity classification. The Linear model consistently outperformed the RBF model across all scenarios, achieving the highest accuracy of 96.54% when balanced with SMOTE. Despite these advancements, challenges persist, including slightly lower recall for minority classes and the potential for overfitting due to synthetic data generation. These findings emphasize the importance of carefully selecting balancing techniques and optimizing hyperparameters to achieve equitable and robust classification performance. This study focuses on SMOTE and RUS for data balancing. Future research could implement hyperparameter tuning, such as Grid CV, and explore other methods that combine oversampling and undersampling for potentially improved results.

ACKNOWLEDGEMENTS

This research was funded by the Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) at Universitas Ahmad Dahlan (UAD) through the Hibah Penelitian Internal Skema Penelitian Dasar for the year 2025. The authors would like to sincerely thank LPPM UAD for their financial support and assistance in making this study possible.

REFERENCES

[1] T. Omer, “The causes of obesity: an in-depth review,” *Adv Obes Weight Manag Control*, vol. 10, no. 4, pp. 90–94, Jul. 2020, doi: 10.15406/aowmc.2020.10.00312.

[2] D. Uğurlu, H. Yapıcı, R. Ünver, and M. Gülü, “Comparison of obesity and physical activity levels of adult individuals by examining dietary habits with different parameters,” *Journal of Health Sciences and Medicine*, vol. 7, no. 3, pp. 301–307, May 2024, doi: 10.32322/jhsm.1450444.

[3] A. De Lorenzo, S. Gratteri, P. Gualtieri, A. Cammarano, P. Bertucci, and L. Di Renzo, “Why primary obesity is a disease?,” *J Transl Med*, vol. 17, no. 1, May 2019, doi: 10.1186/s12967-019-1919-y.

[4] Matthias Blüher, “Obesity: global epidemiology and pathogenesis,” *Nat Rev Endocrinol*, vol. 15, no. 5, pp. 288–298, 2019, doi: 10.1038/s41574-019- 0176-8.

[5] Y. Wang, M. A. Beydoun, J. Min, H. Xue, L. A. Kaminsky, and L. J. Cheskin, “Has the prevalence of overweight, obesity and central obesity levelled off in the United States? Trends, patterns, disparities, and future projections for the obesity epidemic,” *Int J Epidemiol*, vol. 49, no. 3, pp. 810–823, 2021, doi: 10.1093/IJE/DYZ273.

[6] M. Safaei, E. A. Sundararajan, M. Driss, W. Boulila, and A. Shapi’i, “A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity,” Sep. 01, 2021, *Elsevier Ltd*. doi: 10.1016/j.compbiomed.2021.104754.

[7] L. N. Ferreira, L. N. Pereira, M. da Fé Brás, and K. Ilchuk, "Quality of life under the COVID-19 quarantine," *Quality of Life Research*, vol. 30, no. 5, pp. 1389–1405, May 2021, doi: 10.1007/s11136-020-02724-x.

[8] E. Verdú, J. Homs, and P. Boadas-Vaello, "Physiological changes and pathological pain associated with sedentary lifestyle-induced body systems fat accumulation and their modulation by physical exercise," Dec. 01, 2021, *MDPI*. doi: 10.3390/ijerph182413333.

[9] D. Mohajan and H. K. Mohajan, "Obesity and Its Related Diseases: A New Escalating Alarming in Global Health," *Journal of Innovations in Medical Research*, vol. 2, no. 3, pp. 12–23, Mar. 2023, doi: 10.56397/jimr/2023.03.04.

[10] E. A. Silveira, R. R. da S. Filho, M. C. B. Spexoto, F. Haghighatdoost, N. Sarrafzadegan, and C. de Oliveira, "The role of sarcopenic obesity in cancer and cardiovascular disease: A synthesis of the evidence on pathophysiological aspects and clinical implications," May 01, 2021, *MDPI*. doi: 10.3390/ijms22094339.

[11] D. Ryan, S. Barquera, O. Barata Cavalcanti, and J. Ralston, "The global pandemic of overweight and obesity: addressing a twenty- first century multifactorial disease," *In: Haring R, Kickbusch I, Ganten D, Moeti M, eds. Handbook of Global Health. Springer International Publishing:*, pp. 739–773, 2021, doi: 10.1007/978-3-030-05325-3_39-1.

[12] M. A. B. Khan, M. J. Hashim, J. K. King, R. D. Govender, H. Mustafa, and J. Al Kaabi, "Epidemiology of Type 2 diabetes - Global burden of disease and forecasted trends," *J Epidemiol Glob Health*, vol. 10, no. 1, pp. 107–111, Mar. 2020, doi: 10.2991/JEGH.K.191028.001.

[13] Q. An, S. Rahman, J. Zhou, and J. J. Kang, "A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges," May 01, 2023, *MDPI*. doi: 10.3390/s23094178.

[14] M. A. Al-Hashem, A. M. Alqudah, and Q. Qananwah, "Performance Evaluation of Different Machine Learning Classification Algorithms for Disease Diagnosis," *International Journal of E-Health and Medical Communications*, vol. 12, no. 6, 2021, doi: 10.4018/IJEHMC.20211101.oa5.

[15] A. I. Putri *et al.*, "Implementation of K-Nearest Neighbors, Naïve Bayes Classifier, Support Vector Machine and Decision Tree Algorithms for Obesity Risk Prediction," *Public Research Journal of Engineering, Data Technology and Computer Science*, vol. 2, no. 1, pp. 26–33, Apr. 2024, doi: 10.57152/predatecs.v2i1.1110.

[16] S. A. Thamrin, D. S. Arsyad, H. Kuswanto, A. Lawi, and S. Nasir, "Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018," *Front Nutr*, vol. 8, Jun. 2021, doi: 10.3389/fnut.2021.669155.

[17] X. Cheng *et al.*, "Does physical activity predict obesity—a machine learning and statistical method-based analysis," *Int J Environ Res Public Health*, vol. 18, no. 8, Apr. 2021, doi: 10.3390/ijerph18083966.

[18] B. Bonnechère, A. Cuevas-Sierra, J. Jeon, S. Lee, and C. Oh, "Age-specific risk factors for the prediction of obesity using a machine learning approach," *Front Public Health*, vol. 10, 2023, doi: https://doi.org/10.3389/fpubh.2022.998782.

[19] E. Carlos *et al.*, "Machine learning Techniques to Predict Overweight or Obesity," *In CEUR Workshop Proceedings*, vol. 3038, pp. 190–204, 924.

[20] M. Aldraimli *et al.*, "Machine learning prediction of susceptibility to visceral fat associated diseases," *Health Technol (Berl)*, vol. 10, no. 4, pp. 925–944, Jul. 2020, doi: 10.1007/s12553-020-00446-1.

[21] T. Kosolwattana, C. Liu, R. Hu, S. Han, H. Chen, and Y. Lin, "A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare," *BioData Min*, vol. 16, no. 1, Dec. 2023, doi: 10.1186/s13040-023-00330-4.

[22] S. Sreejith, H. Khanna Nehemiah, and A. Kannan, "Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection," *Comput Biol Med*, vol. 126, 2020, doi: https://doi.org/10.1016/j.compbiomed.2020.103991.

[23] E. Ismail, W. Gad, and M. Hashem, "A hybrid Stacking-SMOTE model for optimizing the prediction of autistic genes," *BMC Bioinformatics*, vol. 24, no. 1, Dec. 2023, doi: 10.1186/s12859-023-05501-y.

[24] Wang Lu, "Imbalanced credit risk prediction based on SMOTE and multi-kernel FCM improved by particle swarm optimization," *Appl Soft Comput*, vol. 114, 2022, doi: 10.1016/j.asoc.2021.108153.

[25] P. C. Y. Cheah, Y. Yang, and B. G. Lee, "Enhancing Financial Fraud Detection through Addressing Class Imbalance Using Hybrid SMOTE-GAN Techniques," *International Journal of Financial Studies*, vol. 11, no. 3, Sep. 2023, doi: 10.3390/ijfs11030110.

[26] A. Özdemir, K. Polat, and A. Alhudhaif, "Classification of imbalanced hyperspectral images using SMOTE-based deep learning methods," *Expert Syst Appl*, vol. 178, 2021, doi: https://doi.org/10.1016/j.eswa.2021.114986.

[27] E. Chamseddine, N. Mansouri, M. Soui, and M. Abed, "Handling class imbalance in COVID-19 chest X-ray images classification: Using SMOTE and weighted loss," *Appl Soft Comput*, vol. 129, Nov. 2022, doi: 10.1016/j.asoc.2022.109588.

[28] M. C. Untoro and M. A. N. M. Yusuf, "Evaluate of Random Undersampling Method and Majority Weighted Minority Oversampling Technique in Resolve Imabalanced Dataset," *IT Journal Research and Development*, vol. 8, no. 1, pp. 1–13, Aug. 2023, doi: 10.25299/itjrd.2023.12412.

[29] S. M. Malakouti, M. B. Menhaj, and A. A. Suratgar, "The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction," *Clean Eng Technol*, vol. 15, Aug. 2023, doi: 10.1016/j.clet.2023.100664.

[30] M. Alida and M. Mustikasari, "Rupiah Exchange Prediction of US Dollar Using Linear, Polynomial, and Radial Basis Function Kernel in Support Vector Regression," *Jurnal Online Informatika*, vol. 5, no. 1, pp. 53–60, 2020, doi: 10.15575/join.

[31] R. Mukarramah, D. Atmajaya, and L. B. Ilmawan, "Performance comparison of support vector machine (SVM) with linear kernel and polynomial kernel for multiclass sentiment analysis on twitter," *ILKOM Jurnal Ilmiah*, vol. 13, no. 2, pp. 168–174, Aug. 2021, doi: 10.33096/ilkom.v13i2.851.168-174.

[32] F. Radhiansyah Razak, M. Kunta Biddinika, and H. Yuliasnyah, "Radial Basis Function Model for Obesity Classification Based on Lifestyle and Physical Condition," *Teknologi Informasi dan Komputer*, vol. 192, no. 2, 2024, doi: 10.31961/eltikom.v8i2.1347.