

Improving Obesity Classification Performance through SMOTE and Feature Selection in Machine Learning Models

Vania Rapita Augesta^{1*}, Tsanaiya Rafa Zuhura¹, Septian Eko Prasetyo¹

¹ Program Studi Teknik Elektro, Universitas Negeri Semarang, Indonesia

* Corresponding author: faniabpn@students.unnes.ac.id

Sitasi: V. R. Augesta, T. R. Zuhura, and S. E. Prasetyo, "Improving Obesity Classification Performance through SMOTE and Feature Selection in Machine Learning Models," *Jurnal Teknologi Informasi dan Multimedia*, vol. 8, no. 4, pp. 962–971, 2026.

Received : 13-05-2026
Revised : 03-07-2026
Accepted : 14-09-2026
Published : 24-09-2026

Copyright: © 2026 by the authors. This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Abstract. This study investigates the performance of machine learning models for seven-class obesity classification using Synthetic Minority Over-sampling Technique for Nominal and Continuous features (SMOTE-NC) and model-based feature selection. The UCI obesity dataset initially contained 2,111 records from Colombia, Peru, and Mexico. After removing 24 exact duplicates, 2,087 unique records were retained and divided using a stratified 80:20 training-test split. To prevent data leakage, SMOTE-NC was applied exclusively to the training folds within the cross-validation pipeline. Feature selection was performed using Random Forest-based SelectFromModel with a median importance threshold, reducing 31 encoded features to 16 selected features. Logistic Regression, Support Vector Machine, and Random Forest were evaluated using repeated stratified five-fold cross-validation with ten repetitions. The ablation analysis produced mean macro F1-scores of 93.21% for the baseline, 93.27% for SMOTE-NC only, 93.69% for feature selection only, and 93.89% for the combined SMOTE-NC and feature-selection pipeline. The combined pipeline significantly outperformed the baseline after Holm correction (adjusted $p < 0.001$), whereas SMOTE-NC alone did not produce a significant improvement (adjusted $p = 0.527$). After GridSearchCV optimization, Random Forest achieved 94.98% accuracy and a macro F1-score of 94.84% on the untouched test set. These findings demonstrate that the combined pipeline improves multiclass obesity classification, with feature selection providing the primary performance contribution and SMOTE-NC serving as a complementary class-balancing technique.

Keywords: Obesity Classification; SMOTE-NC; Feature Selection; Random Forest; Multiclass Classification

1. Introduction

Obesity is a complex health condition characterized by excessive body-fat accumulation that may increase the risk of cardiovascular disease, diabetes, musculoskeletal disorders, and other chronic health problems. It differs from overweight, which refers to body weight above the recommended range without necessarily meeting the clinical threshold for obesity. Body Mass Index (BMI), calculated as body weight in kilograms divided by the square of height in meters, is widely used to screen and classify weight status [1, 2]. The World Health Organization defines obesity as having a Body Mass Index (BMI) of at least 30 kg/m^2 and overweight as having a BMI of 25 kg/m^2 or above. The World Health Organization classifies adults with a BMI of 25 kg/m^2 or higher as overweight and those with a BMI of 30 kg/m^2 or higher as obese [1, 3]. Although BMI provides a practical screening measure, it does not fully represent the behavioral, environmental, and hereditary factors associated with obesity [4].

Obesity is influenced by interacting factors such as dietary habits, physical activity, water consumption, technology use, transportation patterns, smoking behavior, alcohol consumption, and family history of overweight. The nonlinear relationships among these variables make machine learning suitable for identifying complex patterns associated with different obesity levels [5, 6]. Machine learning models can process demographic, behavioral, and physical-condition attributes simultaneously, enabling multiclass classification beyond a simple obese and non-obese distinction.

Previous studies have demonstrated the potential of machine learning for obesity classification. One study reported that Random Forest achieved an accuracy of 92.29% and outperformed K-Nearest Neighbors, Naïve Bayes, Decision Tree, and Support Vector Machine in predicting obesity levels [7]. Another study using anthropometric parameters obtained through bioelectrical impedance analysis reported that Random Forest achieved an accuracy of 84.2%, with BMI, Fat Mass Index, and Fat-Free Mass Index identified as influential predictors [8]. These findings indicate that ensemble-based methods can effectively capture nonlinear patterns in obesity-related data.

Nevertheless, several methodological gaps remain in the studies reviewed. First, model comparisons have primarily emphasized aggregate accuracy, providing limited evidence regarding performance across individual obesity classes and the statistical reliability of differences among algorithms. Second, the separate and combined contributions of class-balancing and feature-selection techniques have not been sufficiently isolated through controlled ablation experiments. Consequently, it remains unclear whether reported improvements originate from the classifier, the resampling procedure, feature reduction, or their combination. Third, insufficient reporting of when resampling and feature selection are performed may reduce reproducibility and create a risk of data leakage if information from the test set influences model development.

This study addresses these gaps through a leakage-controlled and reproducible evaluation framework for seven-class obesity classification. The methodological contribution consists of four elements. First, the original data are divided into training and testing subsets before resampling, and SMOTE-NC is applied exclusively within the training folds to accommodate the combination of numerical and categorical predictors. Second, Random Forest-based model selection is used to implement an explicit and verifiable feature-selection procedure. Third, four configurations, consisting of a baseline, SMOTE-NC only, feature selection only, and their combination, are evaluated to quantify the contribution of each technique. Fourth, repeated stratified cross-validation, paired statistical testing with multiple-comparison correction, complete class-level evaluation, and an untouched test set are used to provide a more reliable assessment of model performance.

Accordingly, this study aims to compare Logistic Regression, Support Vector

Machine, and Random Forest for multiclass obesity classification; evaluate the individual and combined effects of SMOTE-NC and model-based feature selection; identify the most influential obesity-related predictors; and determine whether the observed performance differences are statistically significant. This structured evaluation is expected to provide a more transparent and reproducible basis for selecting machine learning methods for obesity classification.

2. Material and Methods

2.1. Dataset and Data Integrity

The Estimation of Obesity Levels Based on Eating Habits and Physical Condition dataset was retrieved directly from the UCI Machine Learning Repository using dataset identifier 544 [6]. The dataset contains 2,111 records, 16 predictor variables, and one target variable named `NObeyesdad`. The target represents seven obesity classes: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. The predictors consist of eight numerical variables, namely Age, Height, Weight, FCVC, NCP, CH2O, FAF, and TUE, and eight categorical variables, namely Gender, `family_history_with_overweight`, FAVC, CAEC, SMOKE, SCC, CALC, and MTRANS. The original dataset contained no missing values. Before model development, 24 exact duplicate records were identified and removed, resulting in 2,087 unique observations. No observations were removed based on the values of Height or Weight.

The unique observations were divided using a stratified 80:20 training-test split with `random_state = 42`. This process produced 1,669 training observations and 418 testing observations while maintaining similar class proportions in both subsets. The test subset was isolated at this stage and was not involved in resampling, preprocessing fitting, feature selection, hyperparameter optimization, or model selection [9].

2.2. Leakage-Controlled Experimental Design

All data-dependent operations were implemented within an imbalanced-learn pipeline to prevent information from the validation or test observations from influencing model training [9]. During cross-validation, each training fold independently fitted the resampling, preprocessing, and feature-selection procedures. The corresponding validation fold was transformed using parameters learned exclusively from its paired training fold [10]. The experimental sequence was therefore defined as follows: duplicate removal, stratified training-test split, cross-validation of the training subset, SMOTE-NC resampling of each training fold, numerical standardization and categorical encoding, model-based feature selection, classifier training, validation-fold prediction, hyperparameter optimization, refitting on the complete training subset, and final evaluation on the untouched test subset.

Four configurations were evaluated to isolate the contributions of resampling and feature selection:

1. Baseline: preprocessing and classification without SMOTE-NC or feature selection.
2. SMOTE-NC only: preprocessing and classification with SMOTE-NC but without feature selection.
3. Feature selection only: preprocessing and classification with feature selection but without SMOTE-NC.
4. Combined configuration: preprocessing and classification with both SMOTE-NC and feature selection.

This ablation design allowed the performance contribution of each technique to be evaluated separately and in combination.

2.3. Preprocessing and SMOTE-NC

SMOTE-NC was selected because the dataset contains both numerical and categorical predictors [11]. Unlike conventional SMOTE, SMOTE-NC distinguishes categorical attributes when identifying neighboring observations and generating synthetic samples. The categorical feature names were explicitly supplied to the sampler. The `sampling_strategy` parameter was set to `auto`, `k_neighbors` was set to 5, and `random_state` was fixed at 42.

SMOTE-NC was applied only to the training portion of each cross-validation fold. Validation observations and the independent test subset were never resampled. When SMOTE-NC was applied to the complete training partition for distribution verification, each of the seven classes contained 281 observations. The training partition increased from 1,669 to 1,967 observations, corresponding to 298 synthetic samples. During cross-validation, however, synthetic samples were generated independently within each training fold.

After resampling, the eight numerical predictors were standardized using `StandardScaler`. The eight categorical predictors were transformed using `OneHotEncoder` with `handle_unknown` set to `ignore`. The preprocessing stage produced 31 encoded predictors. Both `StandardScaler` and `OneHotEncoder` were fitted only using the relevant training-fold observations.

2.4. Model-Based Feature Selection

The correlation matrix was used only as an exploratory analysis of relationships among numerical variables and was not used to remove predictors. The primary feature-selection procedure was implemented using `SelectFromModel` with a `RandomForestClassifier` containing 150 trees and `random_state` = 42 [12]. The selector used impurity-based feature-importance scores and a median threshold. Because the selector was included within the pipeline, feature importance and the selection threshold were recalculated using only the training observations of each cross-validation fold.

After the final pipeline was fitted to the complete training subset, the median importance threshold was 0.016579. Of the 31 encoded predictors, 16 predictors met or exceeded the threshold and were retained, while 15 predictors were eliminated. [Table 1](#) presents the complete feature-selection results.

Table 1. Features retained and eliminated by Random Forest-based `SelectFromModel`

Selected feature	encoded	Importance (%)	Eliminated feature	encoded	Importance (%)
Weight		28.955	MTRANS = Public Transportation		1.480
Age		9.275	MTRANS = Automobile		1.454
Height		8.311	FAVC = No		1.271
FCVC		8.121	FAVC = Yes		1.092
NCP		5.263	SCC = No		0.523
FAF		4.584	CALC = Frequently		0.447
CH2O		4.352	SCC = Yes		0.439
TUE		4.290	CAEC = Always		0.317
Gender = Female		3.452	CAEC = No		0.291
Gender = Male		3.188	MTRANS = Walking		0.261
family_history = No		2.241	SMOKE = Yes		0.198
family_history = Yes		2.186	SMOKE = No		0.177
CALC = Sometimes		2.103	MTRANS = Motorbike		0.0465
CALC = No		2.034	MTRANS = Bike		0.032
CAEC = Sometimes		1.955	CALC = Always		0.004
CAEC = Frequently		1.658	–		–

The selected encoded predictors represented 12 of the 16 original variables. All encoded categories associated with MTRANS, FAVC, SCC, and SMOKE were

eliminated. Weight had the highest selection importance, followed by Age, Height, and FCVC.

2.5. Classification Models and Hyperparameter Optimization

Three multiclass classifiers were evaluated: Logistic Regression, Support Vector Machine, and Random Forest. Logistic Regression was configured with a maximum of 3,000 iterations. The Support Vector Machine used the radial basis function kernel with $C = 1.0$. RandomForestClassifier was used as the underlying estimator for the feature-selection procedure because tree-based models provide impurity-based feature-importance measures that can be used for model-based feature selection [13].

The initial model and ablation comparisons were conducted using stratified five-fold cross-validation with shuffling and `random_state = 42`. Accuracy, macro precision, macro recall, and macro F1-score were recorded. Macro F1-score was treated as the primary metric because it assigns equal importance to each of the seven classes.

After the combined Random Forest configuration produced the highest cross-validation performance, its classifier hyperparameters were optimized using GridSearchCV [14]. The resampling, preprocessing, and feature-selection procedures remained inside the pipeline during every GridSearchCV fold. GridSearchCV evaluated 72 parameter combinations using five-fold stratified cross-validation, resulting in 360 model fits. The estimator was refitted according to the highest macro F1-score.

Tabel 2. Random Forest hyperparameter search space and selected values

Hyperparameter	Values evaluated	Selected value
<code>n_estimators</code>	200, 300, 500	300
<code>max_depth</code>	None, 10, 20	None
<code>min_samples_split</code>	2, 5	2
<code>min_samples_leaf</code>	1, 2	1
<code>max_features</code>	sqrt, log2	sqrt

The selected configuration contained 300 trees, unrestricted maximum depth, sqrt feature sampling, a minimum of two observations to split an internal node, and a minimum of one observation per leaf. This configuration produced a cross-validation macro F1-score of 93.91%, accuracy of 94.13%, macro precision of 94.13%, and macro recall of 93.89%. GridSearchCV used only the training subset. The selected pipeline was subsequently refitted using the complete training subset and evaluated once on the untouched test subset.

2.6. Evaluation Metrics and Statistical Analysis

Model performance was evaluated using accuracy, macro precision, macro recall, macro F1-score, and weighted F1-score [15]. Macro F1-score was selected as the primary evaluation metric because it calculates the F1-score independently for each of the seven target classes and assigns equal weight to every class [15]. The weighted F1-score was additionally reported for the final test evaluation to account for differences in class support.

The stability of the experimental configurations was evaluated using repeated stratified five-fold cross-validation with ten repetitions, producing 50 paired validation scores for each configuration. The same cross-validation splits were used across all configurations and classifiers to ensure paired comparisons.

Statistical differences among the four ablation configurations were first examined using the Friedman test. The Friedman test was used to assess whether statistically significant differences existed among the multiple experimental configurations based on the paired cross-validation results [16]. When the overall test indicated a significant difference, pairwise comparisons were performed using the Wilcoxon signed-rank test on the corresponding 50 paired macro F1-scores.

The Wilcoxon signed-rank test was subsequently used for pairwise comparisons between configurations, following the recommended non-parametric approach for comparing classifier performance across repeated experimental results [16]. The same procedure was applied to compare Random Forest, Support Vector Machine, and Logistic Regression under the combined configuration. Holm correction was applied to control the family-wise error rate across multiple pairwise comparisons [17]. Statistical significance was determined at an adjusted significance level of 0.05. Rank-biserial correlation was reported as the effect-size measure for the Wilcoxon comparisons.

3. Results

3.1. Ablation Analysis of SMOTE-NC and Feature Selection

The contribution of SMOTE-NC and model-based feature selection was examined through four Random Forest configurations. [Table 3](#) presents the macro F1-scores obtained from repeated stratified five-fold cross-validation with ten repetitions.

Table 3. Repeated cross-validation results of the Random Forest ablation study

Configuration	Mean macro F1 (%)	SD (%)	Min (%)	Max (%)
Baseline	93.210	1.336	90.212	95.657
SMOTE-NC only	93.268	1.118	91.065	96.011
Feature selection only	93.690	1.458	90.344	96.905
SMOTE-NC and feature selection	93.886	1.466	90.953	97.228

The combined configuration obtained the highest mean macro F1-score of 93.886%, compared with 93.210% for the baseline, representing an absolute improvement of 0.676 percentage points. Feature selection alone improved the mean macro F1-score by 0.479 percentage points, whereas SMOTE-NC alone produced a smaller improvement of 0.058 percentage points.

The Friedman test identified an overall difference among the four configurations ($\chi^2 = 17.140$, $p = 0.000661$). Therefore, pairwise Wilcoxon signed-rank tests with Holm correction were conducted, as presented in [Table 4](#).

Table 4. Pairwise Wilcoxon signed-rank comparisons among the ablation configurations

Config. A	Config. B	Mean diff. A-B (%)	W	Holm-adj. p	Rank-biserial	Sig.
SMOTE-NC only	Baseline	0.058	571	0.527279	0.104	No
Feature selection only	Baseline	0.479	338	0.013259	0.470	Yes
Combined	Baseline	0.676	242	0.000358	0.620	Yes
Combined	SMOTE-NC only	0.618	230	0.000241	0.639	Yes
Combined	Feature selection only	0.196	422	0.116196	0.311	No
Feature selection only	SMOTE-NC only	0.421	365	0.023533	0.427	Yes

The combined configuration significantly outperformed the baseline and the SMOTE-NC-only configuration. Feature selection alone also significantly outperformed both the baseline and SMOTE-NC alone. However, SMOTE-NC alone did not significantly outperform the baseline, and adding SMOTE-NC to feature selection did not produce a statistically significant improvement over feature selection alone. These findings indicate that feature selection was the primary statistically supported contributor to the observed improvement, while SMOTE-NC provided a small complementary improvement when used in the combined pipeline.

3.2. Comparison of Classification Models

Random Forest, Support Vector Machine, and Logistic Regression were subsequently compared using the combined SMOTE-NC and feature-selection configuration. The results are shown in [Table 5](#).

Tabel 5. Repeated cross-validation performance of the classification models

Model	Mean macro F1 (%)	SD (%)	Min (%)	Max (%)
Random Forest	93.886	1.466	90.953	97.228
Support Vector Machine	91.120	1.473	87.526	95.669
Logistic Regression	87.038	1.964	83.439	91.845

Random Forest achieved the highest mean macro F1-score, exceeding Support Vector Machine by 2.766 percentage points and Logistic Regression by 6.848 percentage points. Pairwise Wilcoxon tests with Holm correction indicated that all model differences were statistically significant.

Tabel 6. Pairwise statistical comparisons among the classification models

Model A	Model B	Mean diff. A–B (%)	W	Holm-adj. p	Rank-biserial
Random Forest	Support Vector Machine	2.766	5	< 0.001	0.992
Random Forest	Logistic Regression	6.848	0	< 0.001	1.000
Support Vector Machine	Logistic Regression	4.082	1	< 0.001	0.998

These results support the selection of Random Forest for hyperparameter optimization and final evaluation.

3.3. Random Forest Hyperparameter Optimization

GridSearchCV evaluated 72 Random Forest hyperparameter combinations through 360 cross-validation fits. The selected classifier used 300 trees, unrestricted maximum depth, sqrt feature sampling, a minimum of two observations required to split an internal node, and one observation per leaf. The selected configuration produced a cross-validation macro F1-score of 93.91%, accuracy of 94.13%, macro precision of 94.13%, and macro recall of 93.89%. The optimized pipeline was subsequently refitted on the complete training subset before being evaluated on the untouched test subset.

3.4. Performance on the Untouched Test Subset

The optimized Random Forest pipeline correctly classified 397 of the 418 test observations. It achieved an accuracy of 94.98%, macro precision of 94.97%, macro recall of 94.79%, macro F1-score of 94.84%, and weighted F1-score of 95.05%. The complete class-wise results are presented in [Tabel 7](#).

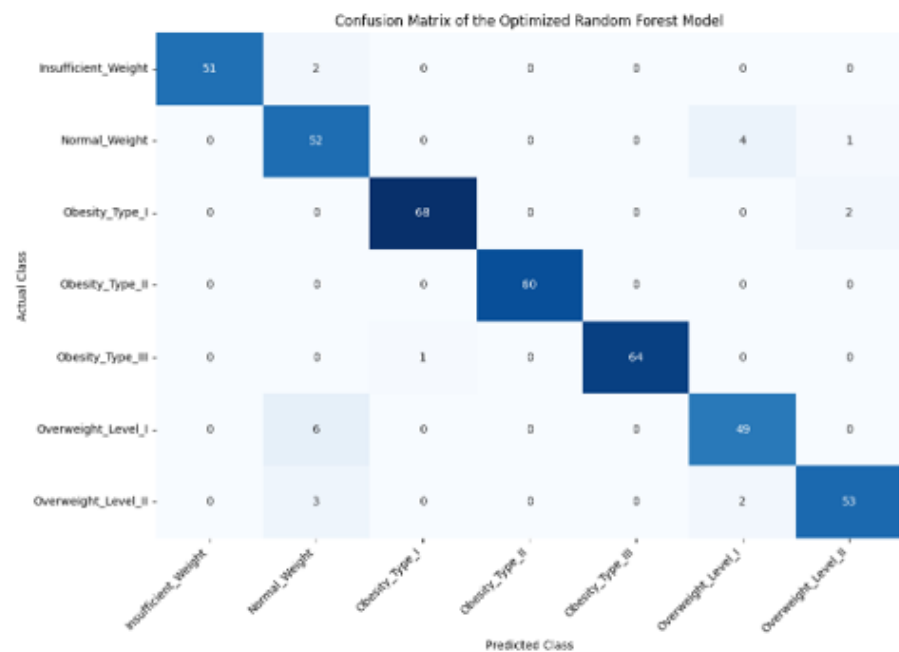
Tabel 7. Classification report of the optimized Random Forest model on the untouched test subset

Class	Precision (%)	Recall (%)	F1-score (%)	Support
Insufficient Weight	100.00	96.23	98.08	53
Normal Weight	82.54	91.23	86.67	57
Obesity Type I	98.55	97.14	97.84	70
Obesity Type II	100.00	100.00	100.00	60
Obesity Type III	100.00	98.46	99.22	65
Overweight Level I	89.09	89.09	89.09	55
Overweight Level II	94.64	91.38	92.98	58
Macro average	94.97	94.79	94.84	418
Weighted average	95.20	94.98	95.05	418

3.5. Selected Features and Feature Importance

After fitting the optimized pipeline to the complete training subset, 16 of the 31 encoded predictors were retained. These predictors represented 12 original variables, while all encoded categories associated with MTRANS, FAVC, SCC, and SMOKE were eliminated.

Based on the aggregated importance scores of the final Random Forest classifier, Weight had the largest contribution at 35.03%, followed by Age at 9.81%,



Gambar 1. Confusion matrix of the optimized Random Forest model on the untouched test subset ($n = 418$)

Height at 9.71%, Gender at 7.90%, and FCVC at 7.78%. The dominance of Weight and Height should be interpreted carefully because the target labels represent weight-status categories that are closely related to body-mass characteristics. Consequently, the importance results describe the predictive behavior of the model and should not be interpreted as evidence of causal relationships.

4. Discussion

The experimental results demonstrate that the leakage-controlled Random Forest pipeline provided consistently higher classification performance than Support Vector Machine and Logistic Regression. Its superiority was observed across 50 paired cross-validation folds and was supported by Wilcoxon signed-rank tests after Holm correction. The optimized model also maintained strong performance on the untouched test subset, obtaining a macro F1-score of 94.84%.

The ablation analysis provides a more specific interpretation of the contributions of SMOTE-NC and feature selection. Although the combined configuration achieved the highest mean macro F1-score, the results do not support the conclusion that SMOTE-NC alone produced a significant performance improvement. SMOTE-NC increased macro F1-score by only 0.058 percentage points relative to the baseline, and this difference was not statistically significant. In contrast, feature selection produced a significant improvement of 0.479 percentage points. Therefore, the main empirically supported improvement was associated with feature selection, while SMOTE-NC served as a complementary mechanism for balancing the training folds.

The relatively small independent contribution of SMOTE-NC may be explained by the moderate initial class imbalance. Before resampling, the training-class counts ranged from 214 to 281 observations, meaning that the minority classes were not extremely underrepresented. Nevertheless, applying SMOTE-NC within the training folds ensured that all classes received balanced representation without exposing validation or test observations to synthetic samples.

The class-wise evaluation showed that Obesity Type II was classified perfectly, while most remaining errors occurred among Normal Weight, Overweight Level

I, and Overweight Level II. These categories represent adjacent weight-status levels and may contain observations with similar physical and behavioral profiles. This overlap helps explain why Normal Weight obtained the lowest class-specific precision and F1-score.

Weight was the most influential predictor in the final model, followed by Age, Height, Gender, and FCVC. However, these importance values reflect predictive associations rather than causal effects. In addition, because the target categories are closely associated with body-weight characteristics, the strong importance of Weight and Height should be considered when interpreting the practical value of the model.

5. Conclusion

This study evaluated the individual and combined contributions of SMOTE-NC and Random Forest-based feature selection to seven-class obesity classification within a leakage-controlled experimental pipeline. Random Forest consistently outperformed Support Vector Machine and Logistic Regression across repeated stratified cross-validation, and the combined SMOTE-NC and feature-selection configuration achieved the highest mean macro F1-score. Statistical testing showed that feature selection was the primary, statistically significant contributor to performance improvement, while SMOTE-NC alone did not produce a significant gain and instead served as a complementary class-balancing mechanism within the combined pipeline. After hyperparameter optimization, the final Random Forest model achieved an accuracy of 94.98% and a macro F1-score of 94.84% on the untouched test subset, with Weight, Age, Height, Gender, and FCVC identified as the most influential predictors. These findings confirm that careful feature selection, rather than resampling alone, provides the more reliable pathway to improved multiclass obesity classification. A limitation of this study is that the dataset originates from a specific regional population (Colombia, Peru, and Mexico), which may limit the generalizability of the model to other populations. Future research may extend this framework to external or more diverse datasets and explore additional feature-selection or ensemble strategies to further improve classification of adjacent weight-status categories.

Artificial Intelligence Use Statement. The authors used Claude.ai for grammar checking and language editing. All output generated by the AI has been checked and verified by the authors. The authors are fully responsible for the accuracy, originality, and integrity of the article's content.

Author Contribution Statement. **VRA** contributed to the conceptualization, methodology, development and implementation of the experimental pipeline, and interpretation of the research results. **TRZ** contributed to data collection and analysis of the research. **SEP** contributed to the drafting, reviewing, and editing of the manuscript. All authors have read, reviewed, and approved the final version of the article and are responsible for all aspects of the research and publication.

Conflict of Interest. The authors declare that there is no conflict of interest related to the research, authorship, and publication of this article.

Funding Statement. This research did not receive any specific funding from government, commercial, or non-profit organizations.

Acknowledgments. The author would like to thank everyone who provided administrative and technical support in the completion of this research.

References

- [1] D. Mohajan and H. K. Mohajan, "Body mass index (bmi) is a popular anthropometric tool to measure obesity among adults," *Journal of Innovations in Medical Research*, vol. 2, no. 4, pp. 25–33, 2023. <https://doi.org/10.56397/jimr/2023.04.06>.
- [2] R. Dieterich, J. Chang, C. Danford, P. W. Scott, C. Wend, and J. Demirci, "The relationship between internalized weight stigma during pregnancy and breastfeeding: A prospective longitudinal study," *Obesity*, vol. 29, no. 5, pp. 919–927, 2021. <https://doi.org/10.1002/oby.23139>.
- [3] A. Ataey, E. Jafarvand, D. Adham, and E. Moradi-Asl, "The relationship between obesity, overweight, and the human development index in world health organization eastern mediterranean region countries," *Journal of Preventive Medicine and Public Health*, vol. 53, no. 2, pp. 98–105, 2020. <https://doi.org/10.3961/jpmph.19.100>.
- [4] OECD, *Health at a Glance: Asia/Pacific 2020*. Health at a Glance: Asia/Pacific, OECD, 2020. <https://doi.org/10.1787/26b007cd-en>.
- [5] A. Chatterjee, M. W. Gerdes, and S. G. Martinez, "Identification of risk factors associated with obesity and overweight—a machine learning overview," *Sensors (Switzerland)*, vol. 20, no. 9, 2020. <https://doi.org/10.3390/s20092734>.
- [6] F. M. Palechor and A. de la H. Manotas, "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico," *Data in Brief*, vol. 25, 2019. <https://doi.org/10.1016/j.dib.2019.104344>.
- [7] H. Syahidah, N. Irsandi, A. N. Ajizah, and A. Amelia, "Obesity prediction using machine learning algorithms," *IJATIS: Indonesian Journal of Applied Technology and Innovation Science*, vol. 2, no. 1, pp. 53–62, 2025. <https://doi.org/10.57152/ijatis.v2i1.1869>.
- [8] R. Yáñez-Sepúlveda *et al.*, "Supervised machine learning algorithms for the classification of obesity levels using anthropometric indices derived from bioelectrical impedance analysis," *Scientific Reports*, vol. 15, no. 1, 2025. <https://doi.org/10.1038/s41598-025-15264-6>.
- [9] Scikit-learn developers, "Common pitfalls and recommended practices." scikit-learn, 2024.
- [10] Scikit-learn developers, "Pipeline." scikit-learn, 2024.
- [11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun 2002. <https://doi.org/10.1613/jair.953>.
- [12] Scikit-learn developers, "Feature selection." scikit-learn documentation, 2024.
- [13] T. Abigail and M. V. Overbeek, "Ensemble learning - random forest algorithm to classify obesity level," *IJNMT (International Journal of New Media Technology)*, vol. 11, no. 2, pp. 68–77, 2025. <https://doi.org/10.31937/ijnmt.v11i2.3709>.
- [14] Scikit-learn developers, "Gridsearchcv." scikit-learn, 2024.
- [15] Scikit-learn developers, "f1_score." scikit-learn, 2024.
- [16] D. Berrar, "Using p-values for the comparison of classifiers: pitfalls and alternatives," *Data Mining and Knowledge Discovery*, vol. 36, no. 3, pp. 1102–1139, 2022. <https://doi.org/10.1007/s10618-022-00828-1>.
- [17] M. A. Valles-Coral *et al.*, "Mathematical modeling and statistical evaluation of hybrid deep learning architectures for multiclass classification of cervical cells in digital papanicolaou images," *Mathematics*, vol. 14, no. 7, p. 1139, 2026. <https://doi.org/10.3390/math14071139>.