DOI: 10.14421/ijid.2026.5714
IJID (International Journal on Informatics for Development), *e, ISSN: 2549-7448*
Vol. 15, No. 1, June 2026, Pp. 21-35

Analyzing Anthropometric Feature Dependency in Obesity Classification Using Feature Ablation

Esti Dwi Puspitasari
Department of Informatics
Amikom University Yogyakarta
Yogyakarta, Indonesia
estidwi@students.amikom.ac.id

Yudha Riwanto*
Department of Informatics
Amikom University Yogyakarta
Yogyakarta Indonesia
yudha.riwanto@amikom.ac.id

Article History
Received November 13th, 2025
Revised March 8th, 2026
Accepted March 13th, 2026
Published July, 2026

Abstract—Obesity classification models often rely heavily on anthropometric features, particularly weight and height, because both variables directly contribute to body mass index (BMI) calculation. This study investigates the extent of such dependency using feature ablation and error analysis. Experiments were conducted on the NObeyesdad dataset using Logistic Regression and Random Forest under three feature configurations: (1) baseline using all features, (2) ablation without weight, and (3) ablation without height. Model performance was evaluated using Balanced Accuracy, Macro F1-score, confusion matrices, and class-level analysis. Results show that Random Forest consistently outperformed Logistic Regression across all scenarios and demonstrated greater robustness to feature removal. Under the baseline setting, Random Forest achieved a Balanced Accuracy of 0.90 and a Macro F1-score of 0.90, while Logistic Regression obtained 0.89 for both metrics. Performance degradation became more pronounced when weight was removed compared with height, indicating stronger model dependence on weight-related information. Error analysis further revealed increased confusion among adjacent obesity categories, particularly overweight classes. These findings highlight the importance of feature dependency evaluation to improve robustness and interpretability in obesity classification models.

Keywords—ablation study; feature dependency; logistic regression; obesity classification; random forest

1 INTRODUCTION

Obesity is a medical condition characterized by the excessive accumulation of body fat, which can cause various serious health problems, such as cardiovascular disease, type 2 diabetes, hypertension, and other metabolic disorders [1], [2]. Factors such as diet, fast food consumption, and a sedentary lifestyle are the main contributors to the rising prevalence of cases of obesity [3], [4]. According to the World Health Organization (WHO) report in 2022, approximately 2.5 billion adults worldwide are overweight, and more than 890 million of them are classified as obese [5], [6]. This figure highlights a complex global health issue that warrants serious attention, both in terms of prevention and early detection.

In contrast to previous studies on obesity classification, which usually focused on achieving high predictive accuracy using complete anthropometric data [7], [8], this study shifts its focus toward analyzing model dependency and feature robustness through ablation studies and error analysis. The existing literature provides limited evidence on systematic comparisons between linear and non-linear models when key anthropometric features are omitted. In particular, it remains unclear whether classification models can still maintain reliable performance when important features related to the body mass index (BMI), such as weight and height, are removed from the input data [9], [10], [11]. An ablation study is therefore used to evaluate the causal impact of feature removal and to examine how model performance changes when specific components are excluded from the learning process [12].

This study aims to assess the level of dependence of the obesity classification model on the anthropometric features, weight, and height, as these variables are the primary components used to calculate the body mass index (BMI), which serves as the main indicator of obesity levels. To achieve this objective, an ablation study approach and structured error analysis are employed to examine how model performance changes when these key features are removed.

Unlike previous studies that primarily aim to introduce new datasets or achieve the highest predictive accuracy, this study focuses on developing a methodological framework to analyze the dependence of obesity classification models on anthropometric features, particularly weight and height. By combining feature ablation experiments with a systematic error evaluation using confusion matrix patterns, this study offers deeper insight into how classification models respond when key anthropometric features are removed. This approach helps determine whether models can grasp indirect relationships between remaining features and anthropometric characteristics when primary measurements are omitted, thereby enabling a structured comparison of how linear and non-linear models react to such feature removal. Thus, the proposed analysis contributes to addressing gaps in existing research by emphasizing model robustness and evaluating feature dependence beyond conventional performance metrics.

2 METHOD

This study presents an experimental framework to assess the extent to which obesity classification models rely on anthropometric features, specifically weight and height features, using an ablation study method.

This study used various test scenarios, including a baseline and several ablation scenarios, to evaluate the performance of the Logistic Regression and Random Forest models accross feature configurations. All experimental stages were designed following a systematic and orderly methodological flow, with the goal of providing fair evaluation between scenarios, avoiding data leakage during the pre-processing and model training stages, and enabling replication the study with a high degree of dependability. The research flow is shown in Figure 1.

2.1 Dataset

The dataset used is NObeyesdad. It includes several features related to the obesity rate of the population in Colombia, Peru, and Mexico aged between 14 and 61 years [13], [14], and consists of seven classes (multi-class).

2.2 EDA

Overall, this dataset consists of 2,111 data points with 17 features [15], [16]. A complete explanation of each feature is provided in Tables 1 and 2.

Some features in the dataset, such as FCVC, NCP, CH2O, FAF, and TUE, are continuous variables synthetically generated by the dataset creator to represent behavioral and lifestyle characteristics. These features were components of the original dataset design and were not generated through oversampling techniques such as SMOTE during this study. In this research, no additional class balancing methods were applied. The dataset was used in its original distribution to preserve the data's natural structure. To address potential class imbalance, evaluation metrics less sensitive to it, such as balanced accuracy and macro F1-score, were used during model evaluation.

2.3 Data Cleaning & Quality Control

This stage aims for accurate analysis and reliable decisions [17]. The following are the steps in data cleaning and quality control:

1. *Missing Value Checking.* Missing values refer to missing data in the dataset [18]. In this study, there are no empty values in the dataset.
2. *Duplicate Data Removal.* Duplicate data was checked using the df.duplicated().sum() function, and 24 rows of duplicate data were detected. Duplicate data was then removed to maintain dataset quality and avoid bias.
3. *Consistency check (Height > 0, Weight > 0).* Checking and removing invalid values, such as height and weight

that are zero or negative, because they are unrealistic. In this dataset, there are almost no zero or negative values.

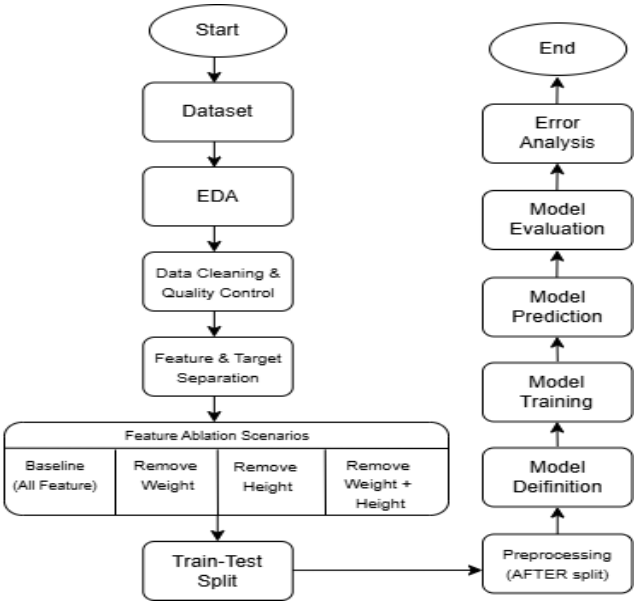


Figure 1. Research workflow with feature ablation scenarios

Table 1. A Table of Features in the Dataset

No	*Features*	*Description*
1	Gender	Gender of the individual (male or female)
2	Age	Age of the individual
3	Height	Height of the individual
4	Weight	Individual body weight
5	family_history_with_overweight	Has a family member suffered or suffers from being overweight?
6	FAVC	Do you eat high, calorie food frequently?
7	FCVC	Do you usually eat vegetables in your meals?
8	NCP	How many main meals do you have daily?
9	CAEC	Do you eat any food between meals?
10	SMOKE	Do you smoke?
11	CH2O	How much water do you drink daily?
12	SCC	Do you monitor the calories you eat daily?
13	FAF	How often do you have physical activity?
14	TUE	How much time do you use technological devices such as cell phones, video games, television, computers, and others?
15	CALC	How often do you drink alcohol?
16	MTRANS	Which transportation do you usually use?
17	NObeyesdad	Obesity level

Table 2. A Table of Features' Data Type

No	*Features*	*Description*
1	Gender	object
2	Age	float64
3	Height	float64
4	Weight	float64
5	family_history_with_overweight	object
6	FAVC	object
7	FCVC	float64
8	NCP	float64
9	CAEC	object
10	SMOKE	object
11	CH2O	float64
12	SCC	object
13	FAF	float64
14	TUE	float64
15	CALC	object
16	MTRANS	object
17	NObeyesdad	object

2.4 Target and Feature Separation

The dataset is divided into input features (*X*). It includes all variables except the obesity label and target (*y*), which is the obesity category (NObeyesdad).

2.5 Train-Test Split

The dataset is separated into training and test data [19][20]. The ratio is 80:20 using stratified sampling based on obesity class.

2.6 Pre-Processing

Pre-processing aims to process raw data into structured information [21], [22]. All steps are combined in the model pipeline via Column Transformer. Pre-processing is tailored to the nature of the model used:

1. *Logistic Regression*, numeric features are normalized using StandardScaler, and categorical features are encoded using One-Hot Encoding.

2. *Random Forest*, numerical features are used without scaling, and categorical features are encoded with One-Hot Encoding.

In StandardScaler, each numeric feature is standardized to have a mean of zero and a standard deviation of one [23]. To prevent data leakage, the scaler is fitted only on the training data after the train-test split. The learned scaling parameters (mean and standard deviation) are then applied to transform both the training and testing datasets. For categorical variables, One-Hot Encoder (OHE) converts each category into binary vectors. Similar to the scaling process, the encoder is fitted on the training data and subsequently applied to the testing data to ensure consistent feature representation while preventing information leakage. Mathematically, categorical data x_i is represented as in (1) [24], [25]:

$$x_i^{(k)} = 1(x_i = c_k), \quad k = 1,2,\dots,K \qquad (1)$$

Where $x_i^{(k)}$ represents the binary encoded value for the $k-$ th category of feature x_i and 1 (.) denotes the indicator function that returns 1 if the condition is satisfied and 0 otherwise.

2.7 Feature Configuration (Ablation Scenarios)

The feature configurations in this study were designed to analyze the extent to which obesity classification models depend on specific input features, particularly anthropometric attributes. To achieve this goal, a feature ablation strategy was implemented, systematically modifying the feature set applied during model training and evaluation. Feature ablation provides an opportunity to observe how model performance changes when specific features are removed from the dataset. Through this approach, the study aimed to identify the strong dependence of classification models on key anthropometric features such as weight and height, and to examine whether the models could retain their predictive ability when these features were partially or completely removed. The feature configurations implemented in this study were divided into several experimental scenarios, including a baseline and several ablation scenarios, which are described in the following subsections.

2.7.1 *Baseline Scenario (All Features):* In baseline scenario 1, all features are used in both model training and testing. This baseline scenario serves as the gold standard for assessing and evaluating model performance before testing for potential feature bias using the ablation approach.

2.7.2 *Ablation Scenario 1 (Without Weight and Height):* In ablation scenario 1, the anthropometric features, weight and height are removed simultaneously from the input feature set. This scenario was created as a direct comparison with the baseline scenario to assess the model's reliance on anthropometric attributes in obesity classification.

2.7.3 *Ablation Scenario 2 (Without Weight):* The partial ablation scenario 2, which does not use the weight feature is designed to isolate the impact of the weight variable on model performance, while retaining the Height feature. This scenario is used to assess whether the model can still significantly differentiate obesity classes when weight information is absent, as well as to observe changes in the model's dependence on features related to behavior and lifestyle.

2.7.4 *Ablation Scenario 3 (Without Height):* A partial ablation scenario 3 without the Height feature was created to evaluate the influence of the Height feature in the obesity classification process, as well as to compare its influence with the removal of the Weight feature in the previous scenario.

2.8 Classification Model

This study uses two different classification approaches, namely Logistic Regression and Random Forest, to represent linear and non-linear modeling strategies. Logistic Regression was chosen because it is a commonly used linear classifier, provides understandable decision boundaries, and serves as a solid foundation for multi-class classification problems. In contrast, Random Forest was chosen as a non-linear ensemble method that can capture complex relationships and interactions between features through multiple decision trees.

The comparison between these two models aims to analyze how different model structures respond to feature removal. Specifically, the linear characteristics of Logistic Regression may indicate a more significant dependence on key features, while Random Forest may exhibit better robustness due to its ability to model non-linear patterns and interactions between features.

This comparison is particularly relevant in obesity classification tasks where both linear relationships (such as anthropometric measurements) and non-linear lifestyle patterns can influence model predictions. Two types of models are applied consistently in the following two scenarios.

2.9 Logistic Regression

Logistic Regression is a statistical method usually used for analysis in health research [26]. It is applied to see the impact of ablation on linear models that are sensitive to variations in feature distributions.

In multi-class classification, Logistic Regression is applied with a multinomial formula [27], where the probability of each class is calculated using the SoftMax function, which converts the linear values of each class into a normalized probability distribution. In the multinomial Logistic Regression model, the Softmax function is used to estimate the probability of each obesity class. With the following probability in (2) [28]:

$$P(y = k|x) = \frac{e^{w_k^T x + b_k}}{\sum_{j=1}^{K} e^{w_j^T x + b_j}} \qquad (2)$$

Where $P(y = k|x)$ represents the probability that the input feature vector x belongs to the class k. The parameter w_k denotes the weight vector for class k_1, b_k represents the bias term, and K is the total number of classes. Then, Table 3 shows the configuration of the Logistic Regression model applied in each experimental scenario.

The configuration parameters applied in Logistic Regression are described in Table 3. Most parameters follow the default settings provided by the implementation to ensure model stability and consistency. The LBFGS solver was chosen because it is suitable for multinomial logistic regression and shows good performance on classification problems involving many classes. The regularization parameter C=1.0 was maintained as the default value to maintain a balance between bias and variance. In addition, the maximum limit for the number of iteration parameters was increased to ensure good training convergence.

2.10 Random Forest

Random Forest is a combined method that utilizes decision trees to produce predictions [29] with the aim of minimizing the relationship between data features [29], [30]. This model is used to assess how well a non-linear model captures patterns from non-anthropometric features and its robustness to the omission of important features. Table 4 shows the Random Forest model configuration settings applied in each experimental scenario.

The Random Forest configurations applied in this study are summarized in Table 4. The number of trees (*n_estimators*) was set at 300 to improve model stability and reduce variance by using more decision trees. The maximum tree depth (*max_depth*) was limited to 8 to regulate model complexity and reduce the possibility of overfitting.

The *min_samples_split* parameter is set to 10 to ensure that internal node splitting is only performed when sufficient training samples are available, which helps generate more general decision boundaries. Additionally, *min_samples_leaf* is set to 5 to avoid creating too few leaf nodes that might capture noise in the dataset. The random_state parameter is set to 42 to ensure reproducible experimental results, while *n_jobs* = -1 allows the algorithm to use all available processor cores to speed up training.

2.11 Model Training and Prediction

2.11.1 Training Process: In the training process, the algorithm learns the patterns that exist between the input variables and the class variables in the dataset [31]. This process is done using the training data set as follows:

- Pre-processing parameters (scaling and encoding) are performed only on training data.
- The classification model was developed with data that had been pre-processed without using test data.
- All training stages are arranged in one workflow (pipeline) to avoid information leakage between stages.

2.11.2 Prediction Process: The prediction process is carried out in the following way:

- The test data is processed using pre-processing parameters that have been learned from the training data.
- The model assigns a predicted label to each observation in the test data.
- At this stage, no parameter adjustments or retraining are performed.

This method ensures that the prediction results reflect the model's generalization ability to previously unencountered data.

2.12 Model Evaluation

Multi-class classification models have more than two classes. Model performance is evaluated using Balanced Accuracy and Macro F1-score, which function to measure average performance across all classes fairly [32], [33].

Table 3. Logistic Regression Configuration Table

LR Configuration Table		
Component	***Parameter***	***Setting***
Model type	-	Logistic Regression
Multi-class strategy	multi_class	Multinomial
Optimization solver	solver	lbfgs
Maximum iterations	max_iter	1000
Regularization	Penalty	12 (default)
Regularization Strength	C	1.0 (default)
Intercept	fit_intercept	True (default)
Pre-processing	-	StandardScaler + OneHotEncoding
Training Framework	-	Pipeline

Table 4. Random Forest Configuration Table

RF Configuration Table		
Parameter	***Selected Value***	***Description***
n_estimators	300	Number of decision trees
max_depth	8	Max depth of each decision tree
min_samples_split	10	Min number of samples to split internal nodes
min_samples_leaf	5	Min number of samples at leaf nodes
random_state	42	Seed for reproducibility
n_jobs	-1	To speed up the training process

2.13 Error Analysis

Error analysis was conducted using a confusion matrix to identify error patterns in the classification of various obesity categories or classes [7]. This analysis aims to provide a deeper understanding of changes in model behavior between the baseline and ablation scenarios.

3 RESULT AND DISCUSSION

The results of this study indicate that body weight is the most influential feature in obesity classification. When this feature is removed, the performance of Logistic Regression significantly decreases from 0.89 to 0.60, indicating that the linear model relies heavily on key anthropometric features for decision-making.

On the other hand, the decrease in Random Forest performance is less drastic (from 0.90 to 0.78), indicating that this model is capable of capturing the complex and non-linear relationships among the remaining behavioral and lifestyle features. This finding suggests that other variables, such as diet and physical activity, may harbor hidden patterns that play a role in obesity classification when processed with a non-linear model.

This finding provides important insight: models for obesity classification may indirectly learn about the relationship between lifestyle features and body composition. Therefore, relying solely on key features such as body weight may obscure the role of other behavioral factors.

The main contribution of this study is demonstrating how feature removal, combined with error analysis, can uncover unseen model dependencies and proxy learning effects in obesity classification.

This study presents a methodological perspective on the evaluation of feature dependencies in machine learning models using feature ablation techniques as well as structured error analysis, offering a deeper understanding of how different model architectures react to the removal of dominant anthropometric features.

3.1 Classification Results with All Features and Ablation Without Weight and Height Features

The experiment was conducted by comparing the baseline model (using all features) and the ablation model (without weight and height features) using two methods, namely Logistic Regression and Random Forest.

3.2 Model Logistic Regression

In the baseline scenario, the model achieved a Balanced Accuracy of 0.895 and a Macro F1-score of 0.893, indicating that Logistic Regression can perform multi-class classification quite effectively when all features are used.

However, after the removal process, the model's performance declined significantly. The Balanced Accuracy value became 0.614, while the Macro F1-score became 0.587. This decline indicates that the model's ability to distinguish between obesity classes becomes significantly more limited when weight and height information are removed. These results are shown in Figures 2 and 3.

To better understand classification performance and error distribution across obesity classes, an error analysis was performed using a confusion matrix. This analysis helps identify which classes are frequently misclassified and reveals patterns in the model's prediction errors.

- *Baseline Logistic Regression*

  Significant misclassification errors occurred in the Normal_Weight and Overweight_Level_I–II classes, where some samples were interchanged. This indicates similarities in characteristics between categories with nearly the same weight range. Figure 4 and Tables 5 and 6 below demonstrate these results.

```
LR Baseline
Balanced Accuracy: 0.8945551639614383
Macro F1: 0.8927601155481121

LR Ablation
Balanced Accuracy: 0.6140300927387381
Macro F1: 0.5865670354490756
```

Figure 2. Logistic Regression model evaluation results

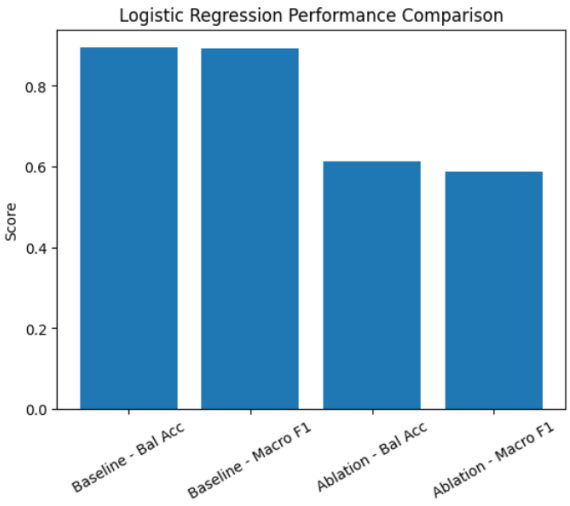


Figure 3. Comparison results of the LR baseline vs ablation evaluation

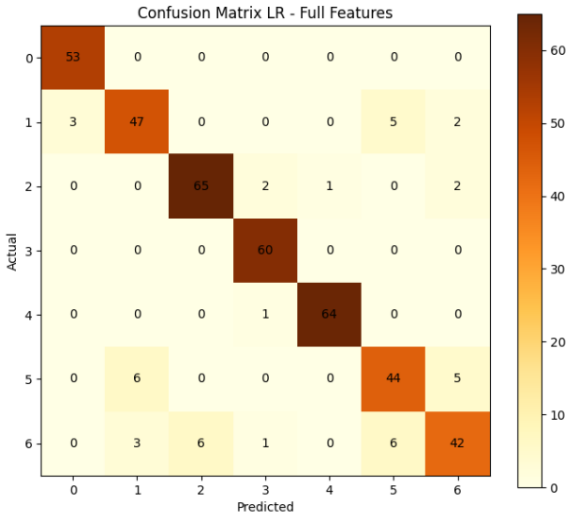


Figure 4. Confusion matrix LR baseline all features

Table 5. Results Table for Each Baseline LR Label

	Table of Results for Each Label			
	TP	***FP***	***FN***	***TN***
Insufficient_Weight	53	3	0	362
Normal_Weight	47	9	10	352
Obesity_Type_I	65	6	5	342
Obesity_Type_II	60	4	0	354
Obesity_Type_III	64	1	1	352
Overweight_Level_I	44	11	11	352
Overweight_Level_II	42	9	16	351

Table 6. Table LR Baseline Classification Report

	Table Classification Report			
	precision	***recall***	***f1-score***	***support***
Insufficient_Weight	0.95	1.00	0.97	53
Normal_Weight	0.84	0.82	0.83	57
Obesity_Type_I	0.92	0.93	0.92	70
Obesity_Type_II	0.94	1.00	0.97	60
Obesity_Type_III	0.98	0.98	0.98	65
Overweight_Level_I	0.80	0.80	0.80	55
Overweight_Level_II	0.82	0.72	0.77	58
accuracy			0.90	418
macro avg	0.89	0.89	0.89	418
weighted avg	0.89	0.90	0.90	418

- *Ablation Logistic Regression (Without Weight and Height Features)*

  After removing the Weight and Height features, classification errors increase significantly. The Confusion Matrix shows that many samples from the Normal_Weight, Overweight_Level_I, and Overweight_Level_II classes were incorrectly predicted as other obesity classes. The Overweight_Level_II class showed the most significant performance decrease, with a recall of only 0.09, indicating that nearly all data in this class was not correctly detected. Figure 5 and Tables 7 and 8 below show these results.

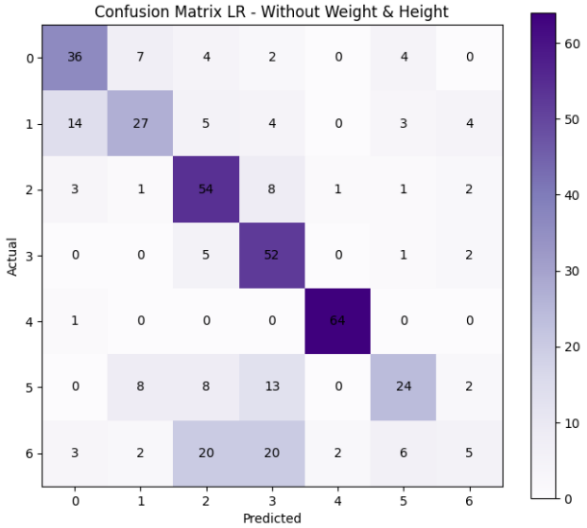


Figure 5. Confusion matrix LR ablation (without weight and height)

Table 7. Results Table for Each Ablation (without Weight and Height Features) LR Label

	Table of Results for Each Label			
	TP	***FP***	***FN***	***TN***
Insufficient_Weight	36	21	17	344
Normal_Weight	27	18	30	343
Obesity_Type_I	54	42	16	306
Obesity_Type_II	52	47	8	311

Obesity_Type_III	64	3	1	350
Overweight_Level_I	24	15	31	348
Overweight_Level_II	5	10	53	350

Table 8. Table LR Ablation Classification Report

	Table Classification Report			
	precision	***recall***	***f1- score***	***support***
Insufficient_Weight	0.63	0.68	0.65	53
Normal_Weight	0.60	0.47	0.53	57
Obesity_Type_I	0.56	0.77	0.65	70
Obesity_Type_II	0.53	0.87	0.65	60
Obesity_Type_III	0.96	0.98	0.97	65
Overweight_Level_I	0.62	0.44	0.51	55
Overweight_Level_II	0.33	0.09	0.14	58
accuracy			0.63	418
macro avg	0.60	0.61	0.59	418
weighted avg	0.61	0.63	0.60	418

3.3 Model Random Forest

In the baseline scenario, the model achieved a Balanced Accuracy of 0.901 and a Macro F1-score of 0.902, indicating excellent performance in identifying classes. In contrast, in the ablation scenario without the weight and height features, the model only achieved a Balanced Accuracy of 0.743 and a Macro F1-score of 0.739. Comparing the results between the two scenarios shows a decrease in Balanced Accuracy of around 15.8% and a decrease in Macro F1-score of around 16.3% after the weight and height features were removed. These results are shown in Figures 6 and 7.

- *Baseline Random Forest*

  Classification results were quite stable across most obesity classes, with high precision, recall, and F1-score values. This indicates that the model can accurately distinguish extreme categories when anthropometric data is available. Figure 8 and Tables 9 and 10 show these results.

```
RF Baseline
Balanced Accuracy: 0.9011560987820765
Macro F1: 0.9021495439344545

RF Ablation
Balanced Accuracy: 0.74308881568532
Macro F1: 0.7393954567432368
```

Figure 6. Random Forest model evaluation results

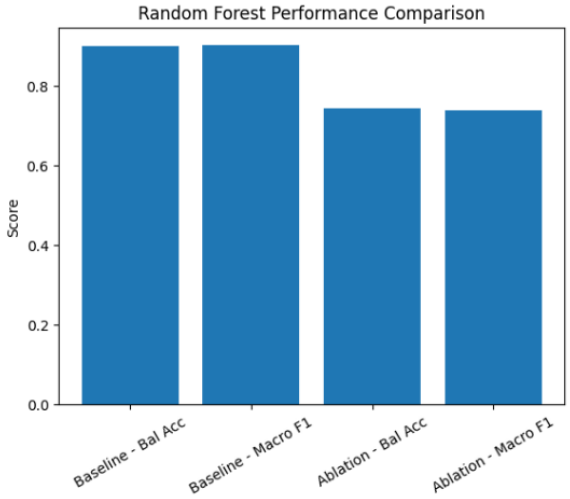


Figure 7. Comparison results of baseline vs ablation evaluation

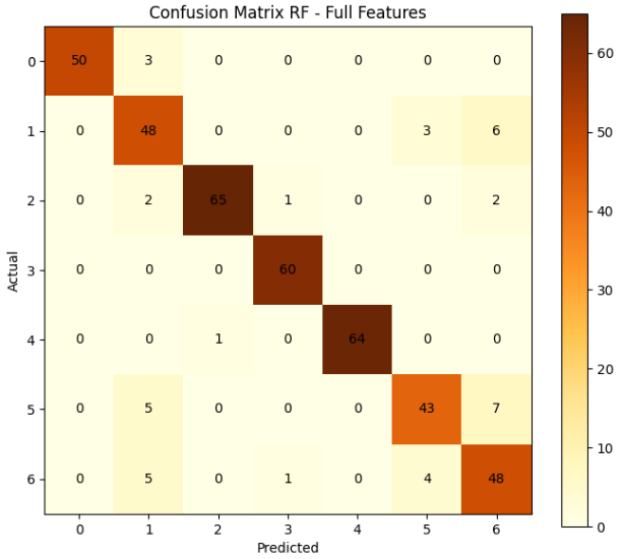


Figure 8. Confusion matrix RF all features

Table 9. Results Table for Each Baseline RF Label

	Table of Results for Each Label			
	TP	***FP***	***FN***	***TN***
Insufficient_Weight	50	0	3	365
Normal_Weight	48	15	9	346
Obesity_Type_I	65	1	5	347
Obesity_Type_II	60	2	0	356
Obesity_Type_III	64	0	1	353
Overweight_Level_I	43	7	12	356
Overweight_Level_II	48	15	10	345

- *Ablation Random Forest (Without Weight and Height Features)*

The performance decline was most pronounced in the Normal Weight, Overweight Level I, and Overweight Level II classes. In the ablation scenario, the recall rate for the Overweight Level II category dropped sharply to 0.41, while the recall rate for the Overweight Level I category dropped to 0.60. This indicates that the model has increased difficulty distinguishing between similar obesity levels when anthropometric data is missing. Figure 9 and Tables 11 and 12 demonstrate these results.

Table 10. Table RF Baseline Classification Report

	Table Classification Report			
	precision	***recall***	***f1, score***	***support***
Insufficient_Weight	1.00	0.94	0.97	53
Normal_Weight	0.76	0.84	0.80	57
Obesity_Type_I	0.98	0.93	0.96	70
Obesity_Type_II	0.97	1.00	0.98	60
Obesity_Type_III	1.00	0.98	0.99	65
Overweight_Level_I	0.86	0.78	0.82	55
Overweight_Level_II	0.76	0.83	0.79	58
accuracy			0.90	418
macro avg	0.91	0.90	0.90	418
weighted avg	0.91	0.90	0.91	418

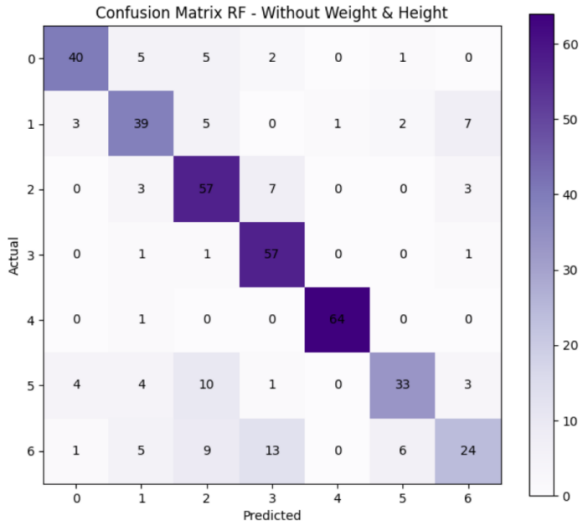


Figure 9. Confusion matrix RF ablation (without weight and height)

The confusion matrix shows that the majority of prediction errors occur between Normal Weight, Overweight Level I, and Overweight Level II. This pattern suggests that without the feature that forms BMI, the boundaries between the classes become less clear.

However, performance in the Type III Obesity category remained quite consistent, with the F1-score remaining around 0.98. These results indicate that obesity classes can still be identified through a combination of behavioral and lifestyle factors, although not as accurately as when anthropometric features are applied. This strengthens the hypothesis that behavioral traits in the severe obesity category tend to be more stable and easier for the model to understand.

3.4 *Classification Results Without Weight Features using Logistic Regression*

In the scenario where the Weight feature was removed, the Logistic Regression model produced a Balanced Accuracy of 0.622 and a Macro F1-score of 0.598. The evaluation results can be seen in Figures 10 and 11.

Based on the class-based assessment results, Obesity Type III performed best with a precision level of 0.96, a recall of 0.98, and an F1-Score of 0.97. This indicates that the majority of data in the extreme obesity class can be identified well, even without weight information.

3.5 *Classification Results without Weight Features using Random Forest*

The Random Forest model achieved a Balanced Accuracy of 0.787 and a Macro F1-score of 0.784. These figures indicate that Random Forest does not rely entirely on the Weight feature in performing multi-class classification for obesity. Compared to Logistic Regression, Random Forest demonstrated more consistent performance.

Table 11. Results Table for Each Ablation RF Label

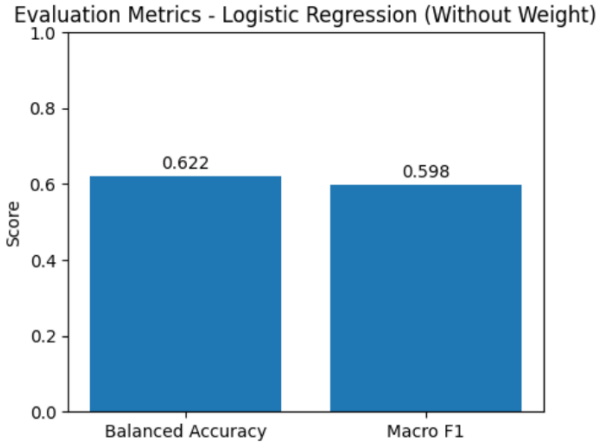
	Table of Results for Each Label			
	TP	***FP***	***FN***	***TN***
Insufficient_Weight	40	8	13	357
Normal_Weight	39	19	18	342
Obesity_Type_I	57	30	13	318
Obesity_Type_II	57	23	3	335
Obesity_Type_III	64	1	1	352
Overweight_Level_I	33	9	22	354
Overweight_Level_II	24	14	34	346

Table 12. Table RF Ablation Classification Report

	Table Classification Report			
	precision	***recall***	***f1-score***	***support***
Insufficient_Weight	0.83	0.75	0.79	53
Normal_Weight	0.67	0.68	0.68	57
Obesity_Type_I	0.66	0.81	0.73	70
Obesity_Type_II	0.71	0.95	0.81	60
Obesity_Type_III	0.98	0.98	0.98	65
Overweight_Level_I	0.79	0.60	0.68	55
Overweight_Level_II	0.63	0.41	0.50	58
accuracy			0.75	418
macro avg	0.75	0.74	0.74	418
weighted avg	0.75	0.75	0.74	418

```
Balanced Accuracy: 0.6223112636463027
Macro F1: 0.5980932078041304
```

Figure 10. LR model evaluation results (without weight features)



Comparison results of LR evaluation without weight

These findings support the results of ablation studies, which showed that Random Forest is more robust to the removal of key features than Logistic Regression. These results can be seen in Figures 13 and 14. The Insufficient Weight, Obesity Type II, and Obesity Type III classes showed quite high levels of prediction, with recall values reaching 0.85, 0.97, and 0.98, respectively. This indicates that extreme classes can still be identified or recognized well even without weight data. Figure 15 and Tables 15 and 16 demonstrate these results.

Figure 11. Confusion Matrix without weight feature

In contrast, the Overweight Level II class recorded the lowest results with a precision of 0.38, a recall of 0.10, and an F1-score of 0.16. Overall, the Macro F1-score of 0.60 indicates that Logistic Regression relies heavily on the Weight feature to distinguish between adjacent obesity classes. Figure 12 and Tables 13 and 14 demonstrate these results.

Table 13. Results Table for Each LR (without Weight Feature) Label

	Table of Results for Each Label			
	TP	***FP***	***FN***	***TN***
Insufficient_Weight	36	20	17	345
Normal_Weight	30	19	27	342
Obesity_Type_I	52	47	18	301
Obesity_Type_II	53	43	7	315
Obesity_Type_III	64	3	1	350
Overweight_Level_I	24	11	31	352
Overweight_Level_II	6	10	52	350

An interesting observation is seen in the Type III Obesity category, where the classification results remain very stable with an F1-score of around 0.98, even when the Weight and Height features are removed. This finding indicates that the Random Forest model can still identify this class by leveraging other attributes in the dataset.

One possible explanation is that individuals in the Type III Obesity category exhibit stronger patterns in lifestyle, related variables such as diet, physical activity levels, or other behavioral indicators. As a non-linear model, Random Forest can capture the complex interactions between these variables, allowing the model to maintain high levels of predictive accuracy even when key anthropometric features are removed.

Table 14. Table LR (without Weight Feature) Classification Report

	Table Classification Report			
	precision	***recall***	***f1-score***	***support***
Insufficient_Weight	0.64	0.68	0.66	53
Normal_Weight	0.61	0.53	0.57	57
Obesity_Type_I	0.53	0.74	0.62	70
Obesity_Type_II	0.55	0.88	0.68	60
Obesity_Type_III	0.96	0.98	0.97	65
Overweight_Level_I	0.69	0.44	0.53	55
Overweight_Level_II	0.38	0.10	0.16	58
accuracy			0.63	418
macro avg	0.62	0.62	0.60	418
weighted avg	0.62	0.63	0.60	418

```
Balanced Accuracy: 0.7871264641161081
Macro F1: 0.7841694861535091
```

Figure 12. RF model evaluation results (without weight features)

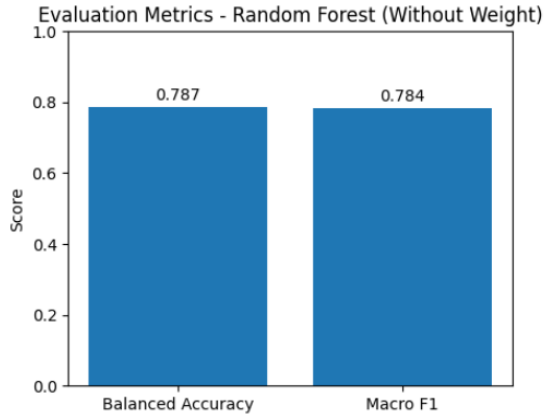


Figure 13. Comparison results of RF evaluation without weight

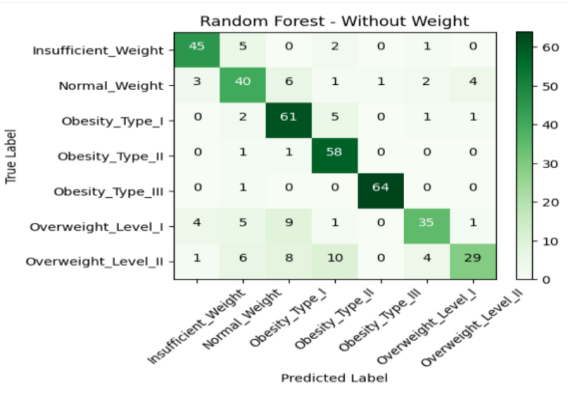


Figure 14. Confusion matrix RF (no weight)

Table 15. Results Table for Each RF (without Weight Feature) Label

ma	Table of Results for Each Label			
	TP	***FP***	***FN***	***TN***
Insufficient_Weight	45	8	8	357
Normal_Weight	40	20	17	341
Obesity_Type_I	61	24	9	324
Obesity_Type_II	58	19	2	339
Obesity_Type_III	64	1	1	352
Overweight_Level_I	35	8	20	355
Overweight_Level_II	29	6	29	354

The results of this study indicate that the severely obese class may have specific habits or physical traits that can still be identified even when key body measurements are removed. This observation further strengthens the hypothesis that non-linear models likely rely more on complex interrelationships between features than on a single dominant anthropometric variable. Table 16 below demonstrates these results.

3.6 Classification Results Without Height Feature using Logistic Regression

This subsection shows the results of the classification model performance assessment, with ablation that does not use the Height feature. The model obtained a Balanced Accuracy of 0.831 and a Macro F1-score of 0.828, which indicates that the ability to classify remains fairly balanced across all classes.

When compared to ablation without the Weight feature, the degradation without the Height feature appears milder, indicating that the Height feature plays a smaller role in the Logistic Regression classification. These results can be seen in Figures 16 and 17.

The highest performance occurred in the Obesity_Type_III and Obesity_Type_II categories, with recall values reaching 0.98 and 0.98, respectively, and F1-scores of 0.99 and 0.94.

On the other hand, the performance decline is more pronounced in the Overweight_Level_II class, which has the lowest recall of 0.53, indicating that nearly half of the samples in this class are still miscategorized into similar classes. Figure 18 and Tables 17 and 18 show these results.

Table 16. Random Forest Table (without Weight feature) Classification Report

	Table Classification Report			
	precision	***recall***	***f1, score***	***support***
Insufficient_Weight	0.85	0.85	0.85	53
Normal_Weight	0.67	0.70	0.68	57
Obesity_Type_I	0.72	0.87	0.79	70
Obesity_Type_II	0.75	0.97	0.85	60
Obesity_Type_III	0.98	0.98	0.98	65

Overweight_Level_I	0.81	0.64	0.71	55
Overweight_Level_II	0.83	0.50	0.62	58
accuracy			0.79	418
macro avg	0.80	0.79	0.78	418
weighted avg	0.80	0.79	0.79	418

```
Balanced Accuracy: 0.8312901236091612
Macro F1: 0.8278340241215292
```

Figure 15. LR model evaluation results (without height feature)

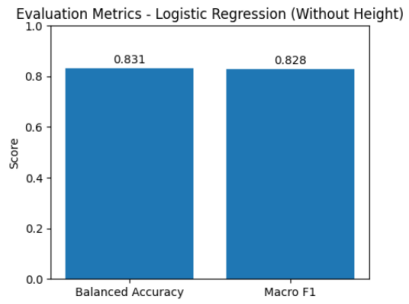


Figure 16. Comparison results of the LR evaluation without height

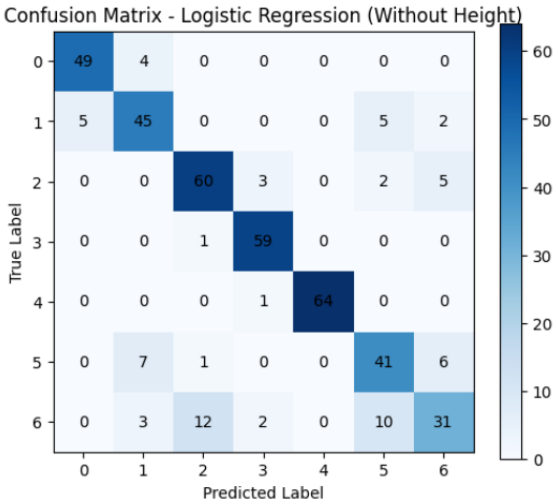


Figure 17. Confusion matrix model LR (without height feature)

Table 17. Results Table for Each LR (without Height Feature) Label

	Table of Results for Each Label			
	TP	***FP***	***FN***	***TN***
Insufficient_Weight	49	5	4	360
Normal_Weight	45	14	12	347
Obesity_Type_I	60	14	10	334
Obesity_Type_II	59	6	1	352
Obesity_Type_III	64	0	1	353
Overweight_Level_I	41	17	14	346
Overweight_Level_II	31	13	27	347

3.7 Classification Results without Height Feature using Random Forest

The Random Forest model still performed well, with a Balanced Accuracy value of 0.87 and a Macro F1-score of 0.87. These figures indicate that the model can maintain equivalent performance across all categories.

Compared to Logistic Regression, Random Forest yields better evaluation scores, indicating a greater ability to capture non-linear relationships among non-anthropometric features. These results indicate that the influence of the Height feature on Random Forest performance is relatively small, and the model can still efficiently distinguish obesity categories using a combination of other features. These results are shown in Figures 19 and 20.

Table 18. Table LR (without Height feature) Classification Report

	Table Classification Report			
	precision	***recall***	***f1-score***	***support***
Insufficient_Weight	0.91	0.92	0.92	53
Normal_Weight	0.76	0.79	0.78	57
Obesity_Type_I	0.81	0.86	0.83	70
Obesity_Type_II	0.91	0.98	0.94	60
Obesity_Type_III	1.00	0.98	0.99	65
Overweight_Level_I	0.71	0.75	0.73	55
Overweight_Level_II	0.70	0.53	0.61	58
accuracy			0.83	418
macro avg	0.83	0.83	0.83	418
weighted avg	0.83	0.83	0.83	418

```
Balanced Accuracy: 0.8732451885014489
Macro F1: 0.873244284515677
```

Figure 18. RF model evaluation results (without height feature)

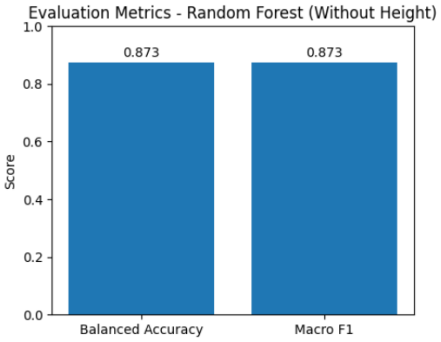


Figure 19. Comparison results of RF evaluation without height

Based on the confusion matrix and per-class analysis, it shows consistent performance with an accuracy of 0.88 and a macro F1-score of 0.87. However, errors occur in adjacent classes, particularly Normal_Weight, Overweight_Level_I, and Overweight_Level_II. The Normal_Weight class has a recall of 0.82, indicating that some samples are still confused with the overweight class. On the other hand, Overweight_Level_II shows the lowest results with a precision of 0.72 and a recall of 0.71, meaning samples in this class are incorrectly classified into neighboring classes such as Overweight_Level_I and Obesity_Type_I. Overall, this error pattern indicates that removing the Height feature affects the separation of the overweight class, but does not significantly reduce the model's ability to detect the extreme obesity class, as seen from the macro F1-score value, which remains high. Figure 21 and Tables 19 and 20 below show these results.

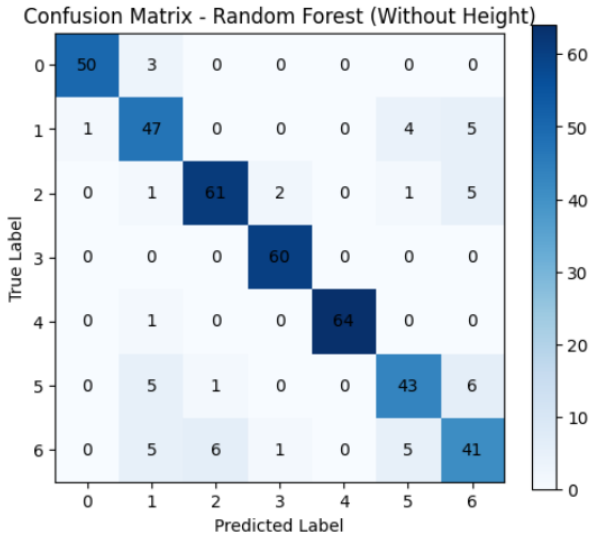


Figure 20. Confusion matrix model RF (without height feature)

Table 19. Results Table for Each LR (without Height Feature) Label

	Table of Results for Each Label			
	TP	***FP***	***FN***	***TN***
Insufficient_Weight	50	1	3	364
Normal_Weight	47	15	10	346
Obesity_Type_I	61	7	9	341
Obesity_Type_II	60	4	0	355
Obesity_Type_III	64	0	1	353
Overweight_Level_I	43	10	12	353
Overweight_Level_II	41	16	17	344

Table 20. Table RF (without Height feature) Classification Report

	Table Classification Report			
	precision	***recall***	***f1-score***	***support***
Insufficient_Weight	0.98	0.94	0.96	53
Normal_Weight	0.76	0.82	0.79	57
Obesity_Type_I	0.90	0.87	0.88	70
Obesity_Type_II	0.95	1.00	0.98	60
Obesity_Type_III	1.00	0.98	0.99	65
Overweight_Level_I	0.81	0.78	0.80	55
Overweight_Level_II	0.72	0.71	0.71	58
accuracy			0.88	418
macro avg	0.87	0.87	0.87	418
weighted avg	0.88	0.88	0.88	418

4 CONCLUSION

This study examines the extent to which obesity classification models are model-dependent and robust to anthropometric features, particularly Weight and Height, using a systematic ablation study using Logistic Regression and Random Forest. The results of the experiment show that the baseline model's high performance is significantly influenced by the presence of Weight and Height features, as evidenced by a marked decrease in performance when both features are removed. Logistic Regression performance experienced a more drastic decline, indicating a strong dependence on the linear relationship between anthropometric features and obesity categories, while Random Forest showed better survival while maintaining fairly good performance using only behavioral and lifestyle features. The Partial Ablation Study also showed that the Weight feature contributed more than the Height feature, so it can be concluded that the Weight feature is the primary source of information indirectly describing BMI. Error analysis shows that most misclassification errors occur within clinically adjacent classes, such as between normal weight and overweight, while the extreme obesity class remains relatively stable even without anthropometric features. This finding confirms that many obesity classification models achieve high performance through the use of BMI proxies, rather than solely based on a comprehensive understanding of behavioral patterns. Thus, this study fills an important gap in the literature by emphasizing the need for robustness evaluation and analysis of model dependability and feature robustness, beyond conventional performance metrics. The proposed ablation-based evaluation framework can serve as a practical approach for assessing feature bias and improving the reliability of machine learning models in healthcare contexts.

CREDIT AUTHOR STATEMENT

Esti Dwi Puspitasari is the first author to conduct a literature review on previous research, data collection, idea research, testing, analysis, and implementation, while Yudha Riwanto, as the second author, provides suggestions and input on the research concept.

COMPETING INTERESTS

In relation to the publication ethics of this journal, Esti Dwi Puspitasari and Yudha Riwanto, the authors of this article, hereby declare that there is no conflict of interest (COI) or competing interest (CI) related to this work.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

Grammarly is used solely as a language editing tool to improve grammar, spelling, and clarity of writing during manuscript preparation. These tools are not used to generate scientific content, formulate research questions, conduct data analysis, interpret results, or draw conclusions. All theoretical concepts, methodological choices, experimental results, and scientific interpretations were developed independently by the authors. The authors have critically reviewed the entire manuscript and are fully responsible for its content.

ACKNOWLEDGMENT

The author would like to express his deepest gratitude to his family for their unwavering support and encouragement throughout this research process. Sincere appreciation is also extended to the lecturers and academic advisors for their guidance, insights, and valuable input in completing this research. In addition, the author would like to thank all parties who have provided assistance and support, both directly and indirectly, in completing this research successfully.

REFERENCES

[1] S. K. Saraswati *et al.*, "Literature Review : Faktor Risiko Penyebab Obesitas," *Media Kesehatan Masyarakat Indonesia*, vol. 20, no. 1, pp. 70–74, 2021, doi: 10.14710/mkmi.20.1.70-74.

[2] B. Busebee, W. Ghusn, L. Cifuentes, and A. Acosta, "Obesity: A Review of Pathophysiology and Classification," *Mayo Clin. Proc.*, vol. 98, no. 12, pp. 1842–1857, 2024, doi: 10.1016/j.mayocp.2023.05.026.Obesity.

[3] T. Islam and M. S. Ali, "Ensemble Learning for Healthcare: A Comparative Analysis of Hybrid Voting and Ensemble Stacking in Obesity Risk Prediction," pp. 1–26, 2025, [Online]. Available: http://arxiv.org/abs/2509.02826

[4] M. Andini, B. Budirman, and B. U. Hasanah, "Faktor Predisposisi, Klasifikasi Subgrup, Biomarker dan Mekanisme Pencegahan Obesitas : Tinjauan Pustaka," *Media Kesehatan Politeknik Kesehatan Makassar*, vol. 20, no. 1, pp. 88–96, 2025, doi: 10.32382/medkes.v20i1.1414.

[5] S. A. Utiarahman and A. M. M. Pratama, "Penerapan Support Vector Machine dan Random Forest Classifier Untuk Klasifikasi Tingkat Obesitas," *Jurnal Fasilkom*, vol. 14, no. 3, pp. 754–760, 2024, doi: 10.37859/jf.v14i3.8104.

[6] N. H. Phelps *et al.*, "Worldwide trends in underweight and obesity from 1990 to 2022: a pooled analysis of 3663 population-representative studies with 222 million children, adolescents, and adults," *The Lancet*, vol. 403, no. 10431, pp. 1027–1050, Mar. 2024, doi: 10.1016/S0140-6736(23)02750-2.

[7] F. R. Valerian, M. Syarief, and D. A. Fatah, "Klasifikasi Tingkat Obesitas Menggunakan Metode Gbm Dan Confusion Matrix," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 2, pp. 2242–2249, 2025, doi: 10.36040/jati.v9i2.13062.

[8] A. K. Tripathi, N. R. Chauhan, and A. Sharma, "Obesity classification and prognosis using machine learning," *AIP Conf. Proc.*, vol. 3224, no. 1, p. 20050, Feb. 2025, doi: 10.1063/5.0245938.

[9] K. A. Wardana and A. M. A. Rahim, "Analisis Perbandingan Algoritma XGBoost Dan Algoritma Random Forest Untuk Klasifikasi Data Kesehatan Mental," *Jurnal Ilmu Komputer dan Pendidikan*, vol. 2, no. 5, pp. 808–818, 2024, [Online]. Available: https://www.kaggle.com/datasets/bhavikjikadara/mental-health-dataset

[10] G. Kasneci and E. Kasneci, "Enriching Tabular Data with Contextual LLM Embeddings: A Comprehensive Ablation Study for Ensemble Classifiers," 2024, [Online]. Available: http://arxiv.org/abs/2411.01645

[11] G. Elkhawaga, O. Elzeki, M. Abuelkheir, and M. Reichert, "Evaluating Explainable Artificial Intelligence Methods Based on Feature Elimination: A Functionality-Grounded Approach," *Electronics (Basel).*, vol. 12, no. 7, p. 1670, Mar. 2023, doi: 10.3390/electronics12071670.

[12] S. Sheikholeslami, M. Meister, T. Wang, A. H. Payberah, V. Vlassov, and J. Dowling, "AutoAblation: Automated Parallel Ablation Studies for Deep Learning," in *Proceedings of the 1st Workshop on Machine Learning and Systems*, New York, NY, USA: ACM, Apr. 2021, pp. 55–61. doi: 10.1145/3437984.3458834.

[13] C. Özkurt, "Examination and Evaluation of Obesity Risk Factors with Explainable Artificial Intelligence," *Computers and Electronics in Medicine*, vol. 1, no. 1, pp. 12–17, 2024, doi: 10.69882/adba.cem.2024072.

[14] Z. Helforoush and H. Sayyad, "Prediction and classification of obesity risk based on a hybrid metaheuristic machine learning approach," *Front. Big Data*, vol. 7, 2024, doi: 10.3389/fdata.2024.1469981.

[15] S. Y. Sibi and A. R. Widiarti, "Klasifikasi Tingkat Obesitas Mempergunakan Algoritma KNN," in *SEMINAR NASIONAL CORISINDO*, Bali, 2022, pp. 370–375. [Online]. Available: https://corisindo.stikom-bali.ac.id/penelitian/index.php/semnas/article/view/73

[16] T. Khater, H. Tawfik, S. Sowdagar, and B. Singh, "Interpretable Models For ML-Based Classification of Obesity," in *Proceedings of the 2023 7th International Conference on Cloud and Big Data Computing*, in ICCBDC '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 40–47. doi: 10.1145/3616131.3616137.

[17] V. Çetİn, "Pamukkale Üniversitesi Mühendislik Bili mleri Dergisi Pamukkale University Journal of Engineering Sciences A comprehensive review on data preprocessing techniques in data analysis Veri analizinde veri ön işleme teknikleri üzerine kapsamlı bir inceleme," vol. 28, no. 2, pp. 299–312, 2022, doi: 10.5505/pajes.2021.62687.

[18] A. Muniasamy, *AI Model Design and Data Management for Disease Prediction*. in Advances in Computational Intelligence and Robotics. IGI Global Scientific Publishing, 2025. doi: 10.4018/979-8-3373-5137-7.

[19] J. M. A. S. Dachi and P. Sitompul, "Comparison Analysis of the XGBoost Algorithm and Random Forest Ensemble Learning Algorithm in Credit Decision Classification," *Jurnal Riset Rumpun Matematika dan Ilmu Pengetahuan Alam*, vol. 2, no. 2, pp. 87–103, 2023, doi: https://doi.org/10.55606/jurrimipa.v2i2.1336.

[20] A. Rahim, I. Pratiwi, and M. Fikri, "Klasifikasi Penyakit Jantung Menggunakan Metode Synthetic Minority Over-Sampling Technique dan Random Forest," *Indonesian Journal of Computer Science*, vol. 12, no. 1, pp. 2995–3011, 2023.

[21] S. S. Al Malaky, A. A. C. Nisa, S. Armiyanti, and R. S. Setyawan, "Comparative Study of Obesity Levels Classification," *JEECS (Journal of Electrical Engineering and Computer Sciences)*, vol. 10, no. 1, pp. 69–75, 2025, doi: 10.54732/jeecs.v10i1.8.

[22] W. Elouataoui, I. El Alaoui, S. El Mendili, and Y. Gahi, "An End-to-End Big Data Deduplication Framework based on Online Continuous Learning," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 9, pp. 281–291, 2022, doi: 10.14569/IJACSA.2022.0130933.

[23] F. Aldi *et al.*, "Standardscaler ' S Potential in Enhancing Breast Cancer," *Journal of Applied Engineering and Technological Science (JAETS)*, vol. 5, no. 1, pp. 401–413, 2023, doi: 10.37385/jaets.v5i1.3080.

[24] E. Poslavskaya and A. Korolev, "Encoding categorical data: Is there yet anything 'hotter' than one-hot encoding?," pp. 92–107, 2023, [Online]. Available: http://arxiv.org/abs/2312.16930

[25] L. Yu, R. Zhou, R. Chen, and K. K. Lai, "Missing Data Preprocessing in Credit Classification: One-Hot Encoding or Imputation?," *Emerging Markets Finance and Trade*, vol. 58, no. 2, pp. 472–482, Jan. 2022, doi: 10.1080/1540496X.2020.1825935.

[26] K. J. Olowe, N. L. Edoh, S. J. C. Zouo, and J. Olamijuwon, "Comprehensive review of logistic regression techniques in predicting health outcomes and trends," *World Journal of Advanced Pharmaceutical and Life Sciences*, vol. 7, no. 2, pp. 016–026, 2024, doi: 10.53346/wjapls.2024.7.2.0039.

[27] A. M. Carreira-Perpiñán and S. S. Hada, "Inverse classification with logistic and softmax classifiers: efficient optimization," *Transactions on Machine Learning Research*, vol. March, no. 2, pp. 1–21, Mar. 2026, [Online]. Available: http://arxiv.org/abs/2309.08945

[28] Y. Yao, J. Zou, and H. Wang, "Optimal Poisson Subsampling for Softmax Regression," *J. Syst. Sci. Complex.*, vol. 36, no. 4, pp. 1609–1625, 2023, doi: 10.1007/s11424-023-1179-z.

[29] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, 2024, doi: 10.58496/BJML/2024/007.

[30] M. Bilal, G. Ali, M. W. Iqbal, M. Anwar, M. S. A. Malik, and R. A. Kadir, "Auto-Prep: Efficient and Automated Data Preprocessing Pipeline," *IEEE Access*, vol. 10, no. October, pp. 107764–107784, 2022, doi: 10.1109/ACCESS.2022.3198662.

[31] M. Kumar and V. Bhardwaj, *Evaluating Label Encoding and Preprocessing Techniques for Breast Cancer Prediction Using Machine Learning Algorithms*, vol. 18, no. 1. Springer Netherlands, 2025. doi: 10.1007/s44196-025-00957-7.

[32] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," *African Journal of Biomedical Research*, vol. 27, no. 4, pp. 4023–4031, 2024, doi: 10.53555/ajbr.v27i4s.4345.

[33] A. Gevaert, A. Jan, R. Thijs, B. Dirk, and V. Tijl, *Evaluating feature attribution methods in the image domain*, vol. 113, no. 9. Springer US, 2024. doi: 10.1007/s10994-024-06550-x.