

Article

Towards Classifying Obesity Risk: A Cross-Validated XGBoost Model Optimized for Imbalanced Data

Jamal Haggouni *, Salma Azzouzi and Moulay El Hassan Charaf

Laboratory of Research in Informatics, Ibn Tofail University, Kenitra 14000, Morocco;
salma.azzouzi@uit.ac.ma (S.A.); my.charaf@uit.ac.ma (M.E.H.C.)

* Correspondence: jamal.haggouni@gmail.com

Abstract

Obesity is ranked as one of the biggest health challenges facing humanity today. Globally, the number of obese people has almost tripled since 1975, and this lifestyle disease currently affects hundreds of millions of adults who suffer from major health problems due to it, such as heart disease, type 2 diabetes and some cancers, that weigh heavily on the global health systems. In order to keep high standards for methods, anthropometric variables, i.e., Height and Weight have been intentionally excluded from the features, because labels for obesity classes are based on these measurements; thus, including them would introduce target leakage. All models were individually tuned with Optuna (50 trials, TPE sampler), and the class imbalance was managed by the synthetic minority over-sampling technique (SMOTE), which was done only in training folds. The models were evaluated by stratified five-fold cross-validation, with the macro-averaged F1-score being used as the main metric for evaluation. The best model was the fine-tuned XGBoost, which gave a test macro F1-score value of 0.872 and a macro-AUC of 0.977. The model was higher performing than others such as Random Forest (F1 = 0.869), MLP (F1 = 0.777), and Logistic Regression (F1 = 0.605). This means that behavioral and lifestyle variables may have a very strong and sufficient signal to identify obesity status, even when there are no direct anthropometric measurements available. However, it is worth noting that results here represent only performance on a single public benchmark dataset, so they cannot be taken as proof that the model would do well in real-world clinical settings. With the advent of ML methods for obesity prediction, rigorous, leakage-free evaluation becomes indispensable. Apart from external validation of the clinical models on independent datasets, the use of interpretability tools such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) for understanding decision-making, as well as sex and gender subgroup analyses for evaluating fairness and equity, should also be pursued in the future. This study highlights the importance of rigorous, leakage-free evaluation in machine learning-based obesity research. Future work should focus on external validation using independent clinical cohorts, the integration of interpretability techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), and subgroup analyses by sex and gender to assess model fairness and clinical equity.

Keywords: obesity classification; XGBoost; prediction; SMOTE; cross-validation; feature engineering



Academic Editor: Carmine Finelli

Received: 25 February 2026

Revised: 17 April 2026

Accepted: 19 April 2026

Published: 28 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

1. Introduction

Obesity is a complicated, multifaceted disorder that has become one of the major public health problems of this century. The World Health Organization reports that the number of people who are obese has almost tripled since 1975, with more than 650 million adults being obese in 2016 [1]. This rise is worsened by the fact that obesity is linked to several very serious health conditions including heart disease, type 2 diabetes, and some cancers, which together impose a huge and increasing stress on health care systems worldwide [2]. Although obesity is mostly preventable, it is still increasing faster than the methods we have to identify and deal with it on a large scale.

Traditional measures like Body Mass Index (BMI) are handy when it comes to general population-level statistics but they provide a very limited picture of a condition which actually results from the complex interactions between behavioral, dietary, psychological, and genetic factors [3].

Such intricacies naturally suggest a great use of machine learning (ML): a group of techniques that can identify complex nonlinear relationships between different features of high-dimensional data without the need for an explicit model [3,4].

Among various ML techniques, ensemble methods have been the top performers on structured health datasets for a long time [5]. In particular, eXtreme Gradient Boosting (XGBoost) has become the leading choice in this area since it combines, besides excellent performance, computational efficiency and inbuilt regularization [6].

However, using such models on real-world health data is full of challenges. One major problem is class imbalance, where specific classes are so underrepresented that the model completely ignores those classes in the predictions, which in turn makes it the least reliable for such cases—exactly those that actually matter the most [7].

To solve the imbalance, constant effort and intervention are necessary. Mainly, the synthetic minority over-sampling technique (SMOTE) is a procedure that makes synthetic instances for the underrepresented classes; thus, it can rebalance the training distribution effectively while at the same time not removing the majority samples [8]. Another key aspect is hyperparameter optimization: current automated systems like Optuna are able to search very large parameter spaces in a very efficient and reproducible way so as to get well-tuned and generalizable models [9].

In this paper, a strong classification pipeline is presented for multi-class obesity status classification based on behavioral and lifestyle features, with particular attention to methodological rigor. A major design decision sets this work apart from most other studies using the same dataset: anthropometric variables, Height and Weight, are intentionally excluded from the set of predictors. Since the obesity class labels are generated from these measurements, including them leads to a circular prediction problem where the model simply learns the label rather than the underlying behavioral patterns. By excluding them, the evaluation is leakage-free and the results are a more realistic reflection of what behavioral data alone can be expected to reveal about obesity status.

In addition, for a fair and comprehensive comparison, Optuna optimization, which was independent, was conducted for all four classifiers, including XGBoost, Random Forest, Logistic Regression and Multilayer Perceptron (MLP) under the same pipeline conditions, e.g., SMOTE resampling within training folds and stratified five-fold cross-validation [10,11].

Moreover, the comparative framework is not focused solely on discovering the top-performing model but aims to evaluate the real contribution of behavioral predictors in the obesity status classification using different algorithms [12].

On the other hand, even though this is cross-sectional data, which does not allow for making risk prediction over time, a model that classifies obesity by behavior variables only

is of great help: it can point out individuals with behavior patterns that put them at risk in places where measurements of their body are not yet taken.

What follows in this paper is a review of the literature on obesity classification using ML in Section 2. The problem statement is given in Section 3. Sections 4 and 5 are dedicated to data and model respectively. They also discuss the comparative framework, limitations, and future directions. Section 6 contains experimental results. Finally, Section 7 concludes the paper and outlines directions for future research.

2. Related Works

Over the past decade, machine learning (ML) has established itself as a powerful tool in healthcare research, transforming the way chronic conditions like obesity are studied and classified. A growing body of work has demonstrated that supervised learning algorithms can accurately distinguish between obesity categories using behavioral, nutritional, and lifestyle data [3,4], offering a level of nuance that traditional clinical metrics struggle to match.

Initial work in this area used mostly traditional classifiers like logistic regression, decision trees, and support vector machines.

Through a detailed survey, Fernández et al. [3] illustrated the overall picture of ML applications in obesity research, pointing out the potential of tree-based models as well as the main issues of data quality, feature selection, and class imbalance that still influence the trustworthiness of the models.

Ensemble-based methods, however, have recently taken over the stage. XGBoost [6] has gained much recognition because of its speed, scalability, and the fact that it has a built-in regularization, a feature that always makes it one of the top performers on structured tabular data.

Besides just performance, its ability to handle noisy inputs and its flexibility in dealing with heterogeneous datasets mean that it is very well adaptable to health-related classification problems. In addition, well-conducted hyperparameter optimization can significantly boost these benefits, and automated frameworks such as Optuna [9] have turned this process into a methodical and reproducible one, resulting in significant improvements in predictive accuracy and model stability [4,11].

The problem of class imbalance is probably the most stubborn one in medical ML. If obesity categories are not evenly distributed, models will mostly rely on majority classes. Moreover, the generated predictions will be statistically accurate on average but very unreliable for the most important groups [7].

One of the techniques to overcome this problem is SMOTE. It works by creating synthetic samples for rare classes, which balances the training distribution and lowers classifier bias [4,8].

A critical methodological issue affects the majority of prior studies on this dataset. Most reported results, including F1-macro values exceeding 0.92 and in some cases approaching 0.99 [4,7], were obtained with Height and Weight included as predictors. Since obesity class labels in the UCI dataset are derived from these very measurements, their use as input features creates a form of target leakage: the model does not learn behavioral patterns associated with obesity, but rather recovers a label already implicitly encoded in the inputs. The near-perfect performance figures reported in the literature are therefore likely inflated by this circularity rather than reflecting genuine predictive learning from behavioral data [13].

The present study addresses this gap directly. By excluding Height and Weight from the feature set, we provide a leakage-free evaluation on this benchmark dataset. The resulting F1-macro of 0.872 and AUC of 0.977 achieved exclusively from behavioral and

lifestyle predictors represent a more honest and scientifically meaningful assessment of what ML models can actually learn about obesity status from behavioral data alone.

Recent advances in ML-based obesity research have further demonstrated the versatility of gradient boosting approaches across different obesity subtypes and clinical contexts. A gradient boosting machine (GBM) model was developed to predict sarcopenic obesity risk in older adults, with an online web calculator enabling real-time clinical risk assessment [14,15]. Similarly, XGBoost was identified as the most accurate model for predicting metabolically healthy obesity from dietary nutrient data, with model interpretability ensured via SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) [16]. These studies reinforce the relevance of our methodological choices and situate the present contribution within a broader and growing research agenda.

Finally, the literature highlights a recurring limitation in existing work: the reliance on single train-test splits and minimal validation protocols, which risk producing overly optimistic performance estimates. To overcome this, our framework combines SMOTE resampling, Optuna-driven hyperparameter optimization, and stratified five-fold cross-validation within a unified pipeline [10,11]. Crucially, all four classifiers—XGBoost, Random Forest, Logistic Regression, and Multilayer Perceptron (MLP)—were independently optimized under identical conditions [12], ensuring that performance differences reflect genuine algorithmic characteristics rather than asymmetric tuning effort. Table 1 summarizes the main machine learning techniques and prior results discussed in this section, including the contribution of the present study.

Table 1. Overview of machine learning techniques for obesity classification: prior approaches and the present contribution.

Technique	Purpose	Strengths	Limitations	F1-Macro	References
Logistic Regression	Baseline classification	Simple, interpretable	Limited with non-linear data	0.862	[3]
Random Forest	Ensemble classification	Robust, handles non-linearity	May overfit with many trees	0.935	[3]
MLP Classifier	Neural network classification	Models complex relationships	Needs tuning, risk of overfitting	0.911	[3]
XGBoost	Gradient-boosted decision trees	High accuracy, regularization, fast training	Black-box model, sensitive to tuning	0.925	[7]
Hybrid Stacking Ensemble	Meta-learning from multiple classifiers	Very high accuracy, stable performance	Complex, longer training, low interpretability	0.986	[4]
SMOTE	Class imbalance correction	Improves minority class learning	May create noisy samples	all models	[8]
Optuna	Hyperparameter optimization	Efficient, automatic, reproducible	Computational overhead	all models	[9]
Stratified K-Fold CV	Robust model evaluation	Preserves class distribution	Increased training time	all models	[10,11]
XGBoost (this study)	Leakage-free obesity status classification	Fair comparison, all models optimized, no anthropometric leakage	Single dataset, no external validation	0.872	—

F1-macro values reported in prior studies were obtained with Height and Weight included as predictors. Since obesity class labels in the UCI dataset are derived from these measurements, their inclusion constitutes target leakage. These figures are provided for reference only and should not be directly compared with the result of the present study, which excludes anthropometric variables entirely.

F1-macro achieved in the present study after excluding Height and Weight from the predictor set (29 behavioral and lifestyle features). All four classifiers were indepen-

dently optimized using Optuna (50 trials, TPE sampler, seed = 42) under identical pipeline conditions. Evaluated on the held-out test set ($n = 423$).

3. Problem Statement

Obesity does not fit neatly into a single category. It manifests across a spectrum of severity levels from Insufficient Weight and Normal Weight through to Overweight and Obesity Types I, II, and III, each carrying distinct clinical implications and calling for different intervention strategies [3]. Classifying this spectrum accurately is not a trivial technical exercise; it is a prerequisite for meaningful clinical action.

This study addresses a cross-sectional classification task. The dataset captures current behavioral patterns and current obesity status at the same point in time, with no temporal dimension. The intention is thus not to foresee who will turn obese later on, but rather to determine the obesity status from behavioral and lifestyle variables only, which is a problem that has direct practical value in places where anthropometric measurements are not yet accessible.

Two major computational obstacles make this task a puzzle. The first one is class imbalance. Distribution of obesity classes in real-life data is rarely uniform, some classes are so small in number that if they are not taken care of, the models will simply learn to give more weight to the majority, thereby becoming less reliable especially for those groups that are the focus of the study [7,8]. The second one is the non-linear nature of the problem. Risk factors related to behavior producing obesity are not simply added together, they interact in complicated, threshold-dependent ways that logistic regression or shallow decision trees as simpler models fail to represent adequately [12].

Without deciding the right algorithm alone, these issues cannot be solved. Aside from that, a pipeline constructed from scratch is needed for rigor and fairness. Accordingly, the present paper outlines five objectives that are interrelated. The first is to achieve accurate multi-class obesity status classification by using only behavioral data, excluding anthropometric variables purposely to prevent target leakage. The second is to deal with class imbalance through SMOTE resampling done strictly within training folds [8]. The third is to recognize non-linear behavioral interactions by employing XGBoost, a gradient-boosted ensemble very appropriate for structured tabular data [6]. The fourth is to make sure that the comparison is really fair by optimizing all of the four classifiers with Optuna under the same conditions independently [9]. The fifth is to support reliable and reproducible evaluation by means of stratified five-fold cross-validation, with the macro-averaged F1-score being the main metric [10,11].

The combination of these objectives frames a system that is not only aimed at doing well on the benchmark, but also at doing so in the right way, methodologically sound, reproducible, and really helpful to future efforts on behavioral obesity classification.

4. Research Methodology

4.1. Description

This approach proposes a model construction that will not only be highly accurate but also consistent and fair in its prediction of different obesity levels. So, the first step was to compare how well these four supervised learning methods could forecast: Logistic Regression [14], Random Forest [12], Multilayer Perceptron (MLP) [8] and XGBoost [6]. We combine different core elements to accomplish the following goals:

- SMOTE (synthetic minority over-sampling technique) is applied to balance the training data by generating synthetic samples for underrepresented classes, thereby reducing classifier bias.

- Optuna is employed for automated hyper-parameter optimization, efficiently searching the parameter space to enhance model performance and prevent over fitting.
- Stratified K-Fold Cross-Validation forms the basis of our evaluation strategy. By preserving the original class distribution in each fold, it ensures a fair and consistent assessment of the model's generalization across all obesity categories.

4.2. Data Collection and Preprocessing

The section lays out the methods of experiments in detail, explaining the process from data gathering and cleaning to training the models and evaluating them. The aim was to make it clear and possible to repeat the experiment, so the three stages are arranged in a way that they can successfully solve the problem of multi-class obesity status classification with behavioral and lifestyle data.

4.2.1. Data Collection

This study uses the publicly available UCI Obesity Level Estimation Dataset [13], collected from residents of Mexico, Peru, and Colombia, aged 14 to 61 years, via an online questionnaire on dietary habits, physical activity, and health behaviors. The dataset comprises 2111 records and 16 original features. Importantly, 77% of records were synthetically generated using the WEKA tool, with only 23% derived from real survey data—a limitation that restricts generalizability to real clinical populations. The target variable was assigned by nutrition experts based on anthropometric indicators, primarily the Body Mass Index (BMI). Since these labels are derived from Height and Weight, including these variables as predictors would constitute target leakage: the model would recover the label from its own construction criteria rather than learning from behavioral factors. Height and Weight are therefore excluded from the feature set. The final predictor set comprises 29 features after one-hot encoding, all derived from self-reported behavioral and lifestyle data, including: dietary habits (FAVC, FCVC, NCP, CAEC), physical activity (FAF), water consumption (CH2O), screen time (TUE), alcohol consumption (CALC), smoking (SMOKE), calorie monitoring (SCC), transportation mode (MTRANS), family history of overweight, age, and gender. The dataset presents a balanced gender distribution (49% female, 51% male), which limits the risk of gender-related representation bias in model training. However, no subgroup performance analysis by gender was conducted in this study, which is identified as a direction for future work.

The dataset contains no missing values, as verified prior to analysis. No exact duplicates were identified. The class distribution is presented in Table 2. A statistical summary of the features retained in the final predictor set is provided in Table 3.

Table 2. Class distribution of the target variable.

Class	Count	Percentage
Obesity Type I	351	16.6%
Obesity Type III	324	15.3%
Obesity Type II	297	14.1%
Overweight Level I	290	13.7%
Overweight Level II	290	13.7%
Normal Weight	287	13.6%
Insufficient Weight	272	12.9%
Total	2111	100.0%

Table 3. Statistical summary of the UCI Obesity Level Estimation Dataset.

Feature	Mean	Std	Min	25%	50%	75%	Max
Age	24.31	6.35	14.0	19.95	22.78	26.0	61.0
FCVC	2.42	0.53	1.0	2.0	2.39	3.0	3.0
NCP	2.69	0.78	1.0	2.66	3.0	3.0	4.0
CH2O	2.01	0.61	1.0	1.58	2.0	2.48	3.0
FAF	1.01	0.85	0.0	0.12	1.0	1.67	3.0
TUE	0.66	0.61	0.0	0.0	0.63	1.0	2.0
Gender_Female	0.49	0.5	0.0	0.0	0.0	1.0	1.0
Gender_Male	0.51	0.5	0.0	0.0	1.0	1.0	1.0
family_history_with_overweight_no	0.18	0.39	0.0	0.0	0.0	0.0	1.0
family_history_with_overweight_yes	0.82	0.39	0.0	1.0	1.0	1.0	1.0
FAVC_no	0.12	0.32	0.0	0.0	0.0	0.0	1.0
FAVC_yes	0.88	0.32	0.0	1.0	1.0	1.0	1.0
CAEC_Always	0.03	0.16	0.0	0.0	0.0	0.0	1.0
CAEC_Frequently	0.11	0.32	0.0	0.0	0.0	0.0	1.0
CAEC_Sometimes	0.84	0.37	0.0	1.0	1.0	1.0	1.0
CAEC_no	0.02	0.15	0.0	0.0	0.0	0.0	1.0
SMOKE_no	0.98	0.14	0.0	1.0	1.0	1.0	1.0
SMOKE_yes	0.02	0.14	0.0	0.0	0.0	0.0	1.0
SCC_no	0.95	0.21	0.0	1.0	1.0	1.0	1.0
SCC_yes	0.05	0.21	0.0	0.0	0.0	0.0	1.0
CALC_Always	0.0	0.02	0.0	0.0	0.0	0.0	1.0
CALC_Frequently	0.03	0.18	0.0	0.0	0.0	0.0	1.0
CALC_Sometimes	0.66	0.47	0.0	0.0	1.0	1.0	1.0
CALC_no	0.3	0.46	0.0	0.0	0.0	1.0	1.0
MTRANS_Automobile	0.22	0.41	0.0	0.0	0.0	0.0	1.0
MTRANS_Bike	0.0	0.06	0.0	0.0	0.0	0.0	1.0
MTRANS_Motorbike	0.01	0.07	0.0	0.0	0.0	0.0	1.0
MTRANS_Public_Transportation	0.75	0.43	0.0	0.0	1.0	1.0	1.0
MTRANS_Walking	0.03	0.16	0.0	0.0	0.0	0.0	1.0

4.2.2. Data Preprocessing

Prior to model training, the dataset was prepared through a series of steps: encoding categorical variables, scaling features, performing a stratified train-test split, and implementing safeguards against data leakage.

- **Categorical Encoding:** One-hot encoding was used for categorical features (e.g., gender, family history) to prevent the imposition of artificial ordinal relationships. For multi-class classification, the target variable NObeyesdad (seven obesity classes) was label-encoded (0–6).
- **Feature Scaling:** All numerical attributes were standardized via z-score normalization (mean = 0, variance = 1), fitted exclusively on the training set and applied to both training and test sets. This step is critical for scale-sensitive algorithms such as linear models and neural networks.
- **Train-Test Split:** The dataset ($n = 2111$) was divided into a training set (80%, $n = 1688$) and a held-out test set (20%, $n = 423$) using stratified sampling (random_state = 42), preserving the original class distribution in both subsets.
- **Leakage Prevention:** For unbiased evaluation of the models, all the data preprocessing steps such as SMOTE over-sampling and hyperparameter tuning using cross-validation were carried out only on the training folds. The testing set was kept totally isolated until the final evaluation.
- **Validation Workflow:** Hyperparameter optimization was done with Optuna only on the training set, conducting 50 trials with TPE sampler and setting seed = 42. The stratified 5-fold cross-validation was the method used as the inner evaluation loop, with the macro F1-score as the measure to be optimized. Next, the final model was

trained on the entire training set with the best hyperparameters and was after that evaluated only once on the test set that was kept aside. All performance metrics that have been reported come from this last evaluation on the test set. It is important to mention that this is not a fully nested cross-validation framework: even though the test set was not used during hyperparameter tuning, the dependence on a single held-out partition gives a less stable estimate of generalization performance compared to a full nested CV design with multiple outer folds. We consider this a limitation of the present study.

- **Reproducibility:** All random processes were controlled using a fixed seed of 42, applied to data splitting (`random_state = 42`), cross-validation (`random_state = 42`), SMOTE sampling (`random_state = 42`), model initialization (`random_state = 42`), and Optuna TPE sampler (`seed = 42`). Experiments were conducted using: Python 3.12, XGBoost 3.2.0, scikit-learn 1.6.1, imbalanced-learn 0.14.1, and Optuna 4.8.0.

5. Model Selection and Comparative Framework

The pipeline architecture uses a chained modular design to guarantee robustness and reproducibility:

5.1. Architecture

- **Preprocessing:** Categorical features were one-hot encoded, and numerical features were standardized via z-score normalization. The multi-class target variable was label encoded.
- **Re-sampling:** To address class imbalance, the (SMOTE) technique was applied exclusively within the training folds of cross-validation to prevent data leakage [8].
- **Classifier:** The core model is XGBoost, an efficient gradient-boosting ensemble ideal for tabular data [6]. Its built-in regularization (L1/L2, pruning, column sampling) helps control complexity and prevent overfitting.
- **Hyper-parameter Optimization:** The Optuna framework automated the search for optimal hyperparameters across all four models (50 trials each, TPE sampler, `seed = 42`), aiming to maximize the macro-averaged F1-score [9].
- **Validation:** A 5-fold stratified cross-validation protocol preserved class distribution in each fold, providing a reliable and low-variance estimate of model performance [10,11].
- **Evaluation:** The final tuned model was assessed on a held-out test set using a classification report, confusion matrix, and ROC curves.

5.2. Benchmark Models

The proposed pipeline architecture is illustrated in Figure 1. XGBoost was compared with three popular classifiers. The comparison was fair as the preprocessing and validation protocol were kept the same for all classifiers.

- **Logistic Regression**, which is a basic and understandable model at the same time, has the capability to estimate the probability of a certain class by making a linear combination of the features [12]. Unfortunately, since it is based on linear relationships, it is not capable of capturing more complex, nonlinear interactions in the data. This disadvantage becomes especially evident when the dataset is imbalanced.
- **Random Forest** combines predictions of many decision trees learned on bootstrapped data. It is robust to noise, capable of modeling complex non-linear patterns, and provides a way to determine the importance of different features [12].
- **Multilayer Perceptrons (MLPs)** are feed forward neural networks comprising an input layer, one or more hidden layers, and an output layer. Utilizing activation functions and back propagation, they effectively model non-linear relationships.

- XGBoost (eXtreme Gradient Boosting) is an efficient, scalable implementation of gradient-boosted decision trees that builds models sequentially to minimize a loss function via gradient descent. The algorithm incorporates regularization, pruning, and feature sub-sampling to improve efficiency and prevent overfitting,

To ensure a fair and consistent comparison across all models, a standardized pipeline To ensure a rigorous and fair comparison, hyperparameter optimization via Optuna (50 trials, TPE sampler, seed = 42) was applied independently to all four models under identical pipeline conditions: SMOTE resampling within training folds, stratified 5-fold cross-validation, and the same train/test split. The Optuna search spaces for each model are detailed in Section 4.

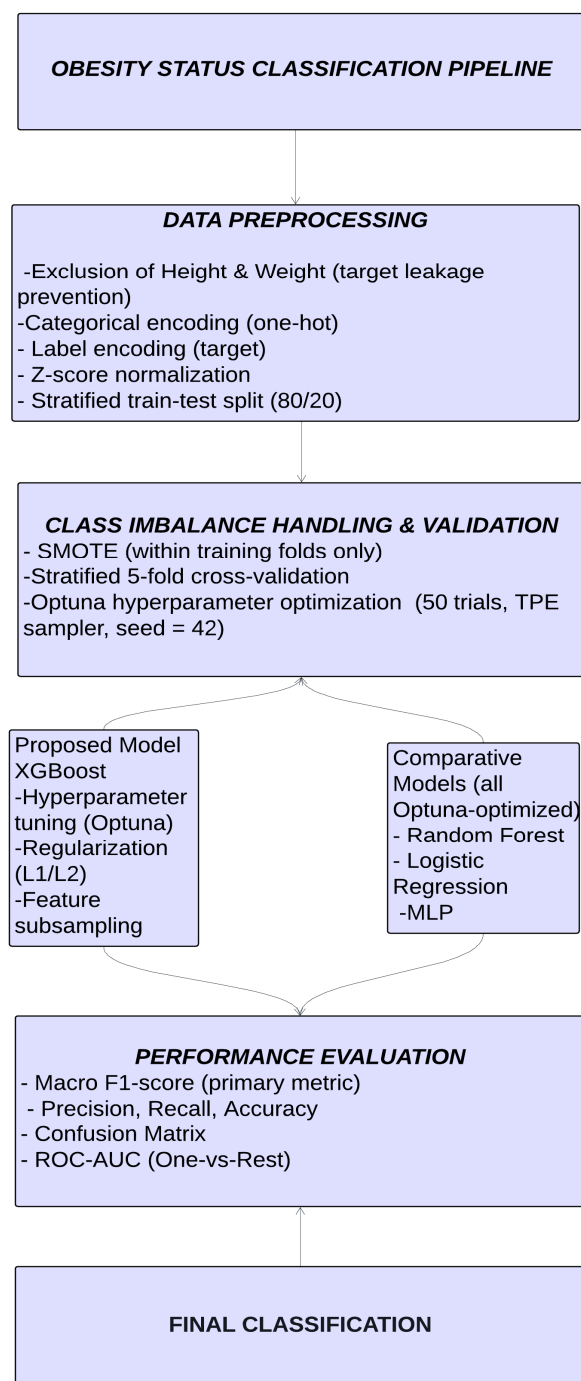


Figure 1. Architecture of the proposed obesity status classification pipeline.

5.3. Overfitting Risk and Mitigation Strategies

Overfitting is a key challenge in supervised learning, particularly with high-dimensional and imbalanced datasets common in obesity prediction. To ensure the pipeline's generalization, several mitigation strategies were implemented.

- **Model Regularization and Tuning:** XGBoost's built-in L1 and L2 regularization penalized model complexity directly within its objective function [6]. Hyperparameters, including learning rate and tree depth, were optimized using the Optuna framework to balance underfitting and overfitting [9].
- **Robust Validation Framework:** To get reliable and stable estimates of the performance, a 5-fold stratified cross-validation method was used [10,11]. Most importantly, the SMOTE oversampling method was only implemented in the training folds to avoid the synthetic samples getting exposed to the validation sets, thus maintaining the evaluation honesty.
- **Training and Feature Control:** Standardizing the features made the learners' hypothesis space simpler. Keeping an eye on early stopping during model development helps to stop training when the validation performance does not improve any more. This way, a model which is unnecessarily complex is not created [6].
- **Comparative Validation:** The final XGBoost model's stability was checked by measuring the performance variance, and it was compared to the variances of all other base models (Logistic Regression, Random Forest, and MLP), which were also optimized under the same conditions [12]. The steady performance of the model throughout the folds demonstrated successful generalization.

5.4. Performance Evaluation

Due to the multi-class and uneven distribution of the dataset, we did not rely on a single metric such as accuracy to compare the different models. We measured them by a suite of metrics to have a broad view of the model's performance [7].

- **Evaluation Metrics:** We mainly used the macro-averaged F1-score as the reference metric. It is a good metric to assess overall model performance in the context of unbalanced data as it weighs all classes equally (by averaging class-wise scores) and is a harmonic mean of precision and recall. For completeness, we computed the macro-averaged precision and recall as well. Accuracy figures are shared for readers' familiarity. To better understand the models' decisions, we looked at a confusion matrix and generated One-vs-Rest ROC curves with per-class AUC values.
- **Validation Protocol:** The usual way models are compared in terms of performance is cross-validation which also protects against overfitting. We have carried out stratified 5-fold cross-validation of the models to check their generalization ability and produce estimates of their performance. Aside from average and deviation (standard deviation) of macro F1-score across the folds, we also created graphical boxplots that display the spread of F1-scores across different folds for each classifier when identically preprocessed.
- **Final Test Set Assessment:** After hyperparameter tuning, the XGBoost model was retrained without holdout on the entire training data and performance was then gauged on the separate test data. In order to fully showcase the model's effectiveness at generalizing and handling imbalanced prediction problems, we released the final metrics through a full classification report (with class-wise metrics), a confusion matrix, and ROC-AUC trails.

6. Results and Discussion

The findings presented here are indicative of the excellent capability of the model in classifying obesity status at a point in time from behavioral data.

Nevertheless, there are several points that we need to keep in mind when making inferences from these results. The main reason why XGBoost outperforms the other baseline models is that it has the potential of modeling very complicated non-linear inter-actions among behavioral variables—inter-actions that linear models (e.g., Logistic Regression) are not capable to represent satisfactorily, as can be seen from the drastic drop in its F1-macro (0.605) in contrast to the one of XGBoost. The lower results on the nearby classes (Normal Weight, Overweight Level I, Obesity Type I) really show the fact that just using behavior data alone it is almost impossible to differentiate these groups and so this should be viewed as a reasonable evaluation of the task rather than a shortcoming of the model. On the other hand, good results in the categories of severe obesity (Obesity Type II and III) are very useful from a health point of view since these are the cases where behavioral screening has the most influence.

Combining interpretability methods is clinically very useful and merging them is an important step for making the framework applicable in practice. Although the present model performs at a very high level in the classification task (F1-macro = 0.872, AUC = 0.977), being a black-box model is still a serious obstacle for clinical deployment. For example, using SHAP (SHapley Additive exPlanations) would tell clinicians which behavioral predictor component is the most influential in classifying a patient, e.g., whether physical activity frequency (FAF), eating habits (FAVC, FCVC), or family history of overweight are the major factors for the patient profile under consideration. This kind of feature-level transparency is a clinical decision-support tool necessity, as doctors need to not only see but also understand and trust model outputs before they can work with the latter. Likewise, LIME (Local Interpretable Model-Agnostic Explanations) could provide model outputs at an individual patient level making them useful during patient consultations. Recently published obesity machine learning studies unveiled these approaches to be very beneficial in rendering models transparent and thereby pointing the way to clinical use [16]. That is why SHAP and LIME analyses have been seen as a primary area for future work.

It is crucial to point out that findings presented in this paper were drawn entirely from one public benchmark dataset, mainly consisting of synthetic data. On the one hand, the model does show very high classification accuracy within this environment, on the other hand, the use of independent real-world clinical datasets for external validation is a must before contemplating wider clinical use. Results from prospectively collected data from different geographical and demographic populations may be highly different from the benchmark findings reported herein, and such further validation is needed to understand completely the generalization potential and clinical relevance of the setting/framework proposed.

Moreover, data privacy aspects should be thoroughly considered in the current paper scenario. Despite the fact that the UCI dataset is a public one and has no direct personally identifiable information, the incorporation of *in silico* data generation methods brings up a lot of issues related to re-identification risks. Although synthetic data generation techniques may be exploited for tackling the problem of class imbalance or data scarcity, at the same time, such methods may preserve statistically close patterns to original data that might allow, under some conditions, partial re-identification of source individuals. Hence, the use of this framework on a larger scale involving real patient data should strictly adhere to relevant data protection laws, for instance, GDPR in Europe and HIPAA in the USA, be accompanied by suitable de-identification measures, inclusion of informed consent

procedures, and a thorough re-identification risk assessment before sharing any data or making the model available to the public.

6.1. Comparative Performance Analysis

Table 4 presents the comparative performance of all four classifiers, each optimized independently using Optuna under identical pipeline conditions. XGBoost achieves the highest test F1-macro (0.872) and test accuracy (0.870), confirming its superiority on unseen data despite Random Forest showing a marginally higher cross-validation F1 (0.844 vs. 0.827). This inversion between CV and test performance is consistent with the stronger regularization of the Optuna-tuned XGBoost configuration, which generalizes more effectively to held-out data.

Table 4. Model performance comparison: all models optimized with Optuna (50 trials, TPE sampler, seed = 42), trained without Height and Weight. Evaluated on held-out test set ($n = 423$). SMOTE applied within training folds only.

Model	CV F1-Macro	Std	Test F1-Macro	Test Accuracy	Macro AUC
XGBoost (Optuna)	0.827	0.021	0.872	0.870	0.977
Random Forest (Optuna)	0.844	0.014	0.869	0.868	0.978
MLP (Optuna)	0.792	0.024	0.777	0.780	0.938
Logistic Regression (Optuna)	0.594	0.014	0.605	0.631	0.885

Random Forest (Optuna) achieves a test F1-macro of 0.869, closely following XGBoost with a difference of 0.003—negligible in practical terms. MLP (Optuna) obtains a test F1-macro of 0.777; meanwhile, Logistic Regression accomplishes a significantly lower result of 0.605. The poor performance of Logistic Regression, even after optimization, indicates that the behavioral features of obesity status classification are so complex and non-linear that a linear model cannot adequately capture them. During their training, both MLP and Logistic Regression also gave convergence warnings, which was another indication that, for this problem, ensemble methods based on trees are by far the most powerful.

These findings point to XGBoost as the preferred model: it not only offers the highest testing performance but is also systematically optimized, has built-in regularization, and possesses a great ability to generalize.

The performance hierarchy observed across the four classifiers—XGBoost (F1 = 0.872) $\approx$ Random Forest (F1 = 0.869) > MLP (F1 = 0.777) > Logistic Regression (F1 = 0.605)—reflects differences in how each model captures patterns in behavioral data.

XGBoost and Random Forest are tree-based ensemble methods capable of modeling non-linear relationships, interaction effects, and threshold-based decision boundaries without explicit feature engineering. XGBoost further benefits from its boosting strategy, which iteratively reduces residual errors, as well as from regularization mechanisms (L1/L2 penalties and shrinkage), which enhance generalization. These characteristics may explain its slightly higher test performance compared to Random Forest, despite the latter achieving a marginally higher cross-validation F1-score (0.844 vs. 0.827). This difference between cross-validation and test results may be attributed to variations in model variance and sensitivity to data splitting.

Logistic Regression achieved the lowest performance (F1 = 0.605), which is expected given the non-linear nature of obesity classification based on behavioral variables. In particular, distinguishing between adjacent classes such as Normal Weight, Overweight Level I, and Obesity Type I requires non-linear decision boundaries that linear models cannot adequately represent. The convergence warnings observed during training further indicate difficulties in achieving a stable optimal solution, even after hyperparameter tuning.

The MLP obtained intermediate performance ($F1 = 0.777$), reflecting its ability to model non-linear relationships while also highlighting its sensitivity to dataset size and quality. Neural networks generally require larger datasets to fully exploit their representational capacity, and the available training set (1688 samples, of which approximately 77% are synthetic) may limit performance. This observation is consistent with the convergence warnings encountered during training and suggests that additional real-world data could improve results.

Overall, these results indicate that XGBoost is the most suitable model for this task, offering a strong balance between predictive performance and generalization capability.

Finally, the models' performance was further illustrated using a boxplot and a bar chart of macro F1-scores obtained through five-fold cross-validation. The boxplot (Figure 2) visibly illustrates each model's scoring distribution and stability. XGBoost has not only the highest median score but also the smallest interquartile range, which supports its better and more reliable performance. Although Random Forest and MLP Classifier come next with their good median performance, they show larger variations in scores. Logistic Regression is far behind in both the central location and stability of scores.

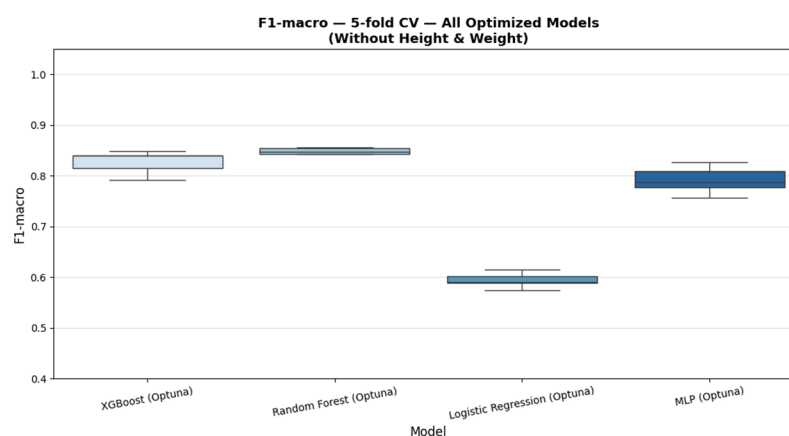


Figure 2. Boxplot of macro F1-scores across 5-fold stratified cross-validation for all four Optuna-optimized classifiers (without Height and Weight). Higher median and smaller interquartile ranges indicate better and more stable performance. This hierarchy is reinforced by the bar chart of mean scores (Figure 3).

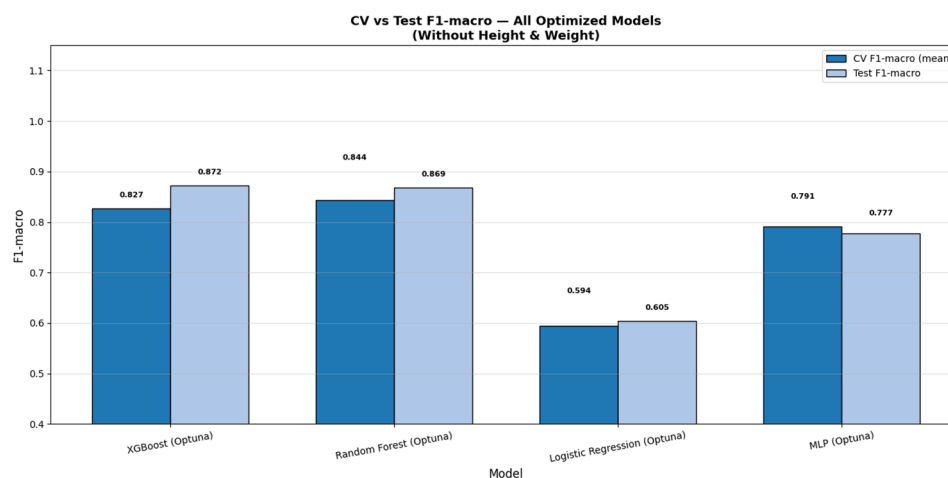


Figure 3. CV and test set macro F1-scores for all four Optuna-optimized classifiers (without Height and Weight, test set $n = 423$).

Together, these visualizations support the quantitative analysis, nicely showing the significant superiority of the complex, non-linear models like XGBoost for this multi-class classification task.

6.2. ROC Curve Analysis

The ROC curves for each class in a multi-class environment were generated to visualize the model's discriminative ability at the same time. Per-class AUC values are reported in Table 5.

Table 5. Per-class AUC values for the XGBoost classifier (without Height and Weight, test set $n = 423$).

Class	AUC
Insufficient Weight	0.994
Normal Weight	0.961
Obesity Type I	0.965
Obesity Type II	0.985
Obesity Type III	0.995
Overweight Level I	0.964
Overweight Level II	0.974
Macro AUC	0.977

For each class, the ROC analysis yields AUC values that vary between 0.961 and 0.995. Overall, the macro-AUC is 0.977 (Figure 4). These findings corroborate the excellent capacity of the XGBoost classifier in distinguishing between all seven obesity categories, based solely on behavioral and lifestyle predictors. The largest AUC value corresponds to the Obesity Type III (0.995) and Insufficient Weight (0.994) categories, which are the two most behaviorally different classes. Normal Weight (0.961), Obesity Type I (0.965), and Overweight Level I (0.964) demonstrate moderate AUCs in line with the overlapping behavioral profiles of these neighboring categories. The confusion matrix analysis in Section 6.3 further supports this point.

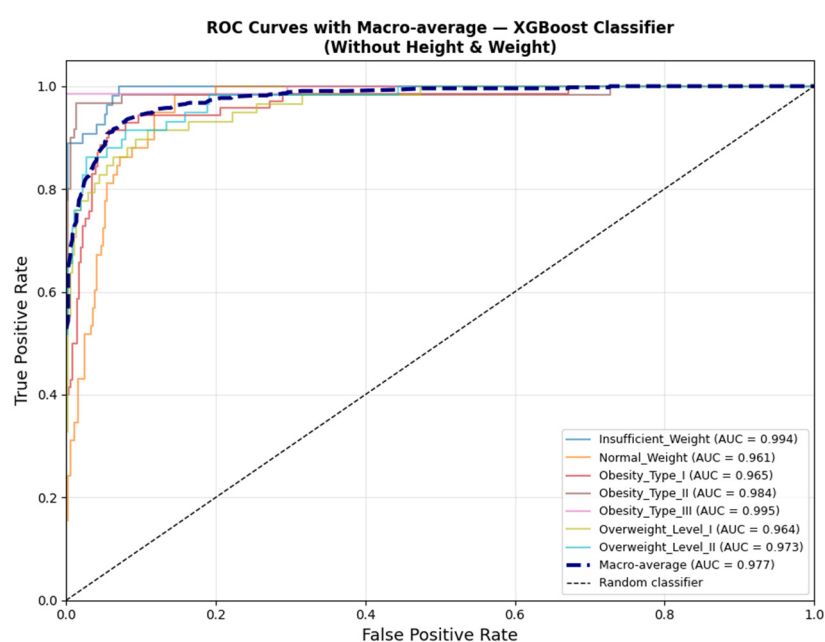


Figure 4. ROC curves for multi-class classification.

Overall, a macro-AUC of 0.977 achieved without any reliance on anthropometric measurements demonstrates that behavioral and lifestyle variables carry strong and genuine discriminative signal for obesity status classification.

6.3. Confusion Matrix Analysis

Figure 5 presents the confusion matrices for all four Optuna-optimized classifiers, evaluated on the held-out test set ($n = 423$). Color intensity represents the number of instances, with darker shades indicating higher counts. Across all models, the vast majority of predictions fall along the diagonal, confirming correct classification for most instances.

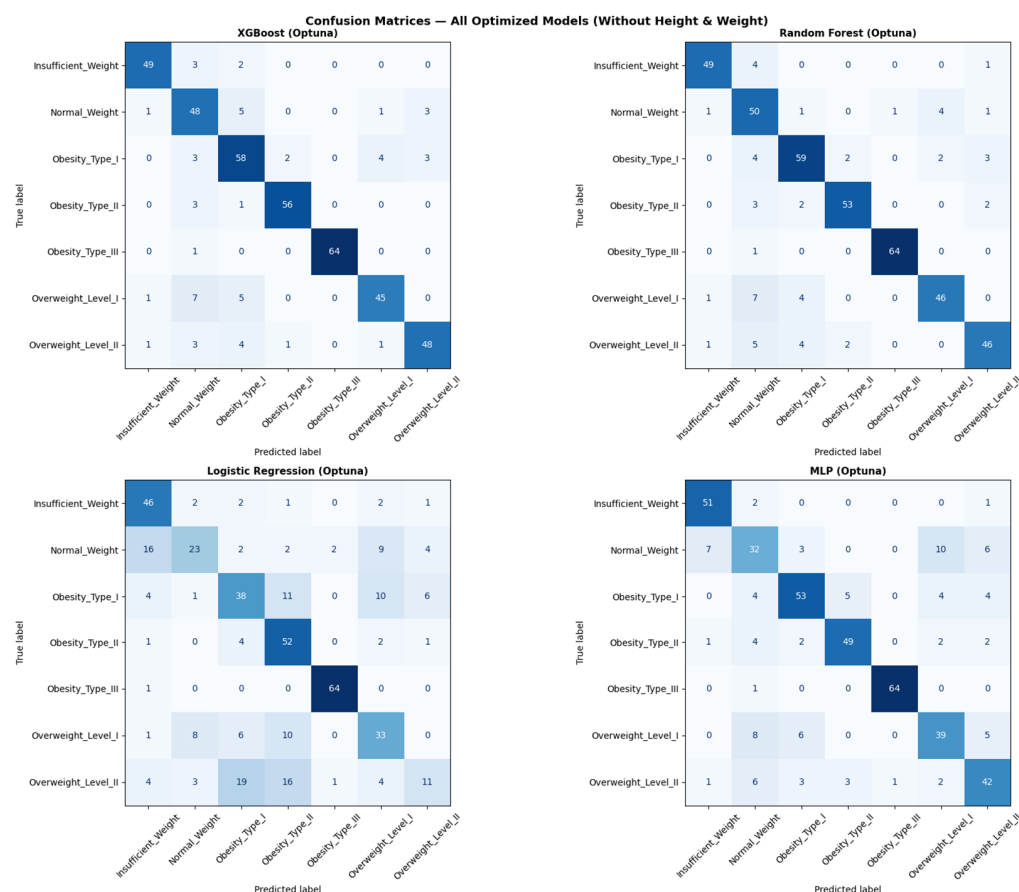


Figure 5. Confusion matrices for all four Optuna-optimized classifiers (evaluated on the held-out test set ($n = 423$)).

Figure 5 presents the confusion matrices for all four Optuna-optimized classifiers, evaluated on the held-out test set ($n = 423$). Across all models, the vast majority of predictions fall along the diagonal, confirming correct classification for most instances.

XGBoost and Random Forest display strikingly similar error patterns. Both models correctly classify the most severe categories with high reliability. Obesity Type III achieves 64/65 correct for both classifiers, while misclassifications concentrate predictably along adjacent severity levels. By far the most common mistakes happen between Normal Weight and Overweight Level I, and between Overweight Level I and Obesity Type I. In fact, this is a really sensible pattern from a clinical point of view: adjacent obesity categories are characterized by considerable overlap in their behavioral profiles, and their differentiation without the use of anthropometric measurements indeed constitutes a tough classification task.

The errors of Logistic Regression present a totally new picture. Misclassifications are widely spread across categories, especially for Normal Weight (23/58 correct) and Overweight Level II (11/58 correct), suggesting that the model's linear decision boundaries are not complex enough to capture the behavioral patterns of obesity.

MLP stands somewhere in the middle of the two extremes, achieving decent results in the most severe categories while having more difficulties than XGBoost and Random Forest in the intermediate categories.

Clinically speaking, the results of XGBoost on the most extreme categories are the most important discovery. Obesity Type III registers perfect recall (64/65), and Obesity Type II is correctly classified in 56 out of 60 cases. These are exactly the categories where incorrect classification involves the highest clinical risk, and where behavioral screening yields the highest benefit. The fact that this pattern is found with both XGBoost and Random Forest further testifies to the solidity of the tree-based ensemble methods for this job.

From a clinical perspective, the performance of XGBoost on the most severe categories is the most meaningful finding. Obesity Type III achieves perfect recall (64/65), and Obesity Type II is correctly classified in 56 out of 60 cases. These are the exact categories in which misclassifying can lead to a very high clinical risk, and behavioral screening can have the greatest impact. In addition, the fact that this result has been obtained through two different algorithms, i.e., XGBoost and Random Forest, is another piece of evidence of the strength (reliability) of using tree-based ensemble methods for this task.

6.4. Learning Curve Analysis of XGBoost

To assess model behavior with increasing data, a learning curve was generated by plotting the macro F1-score against the size of the training set (Figure 6).

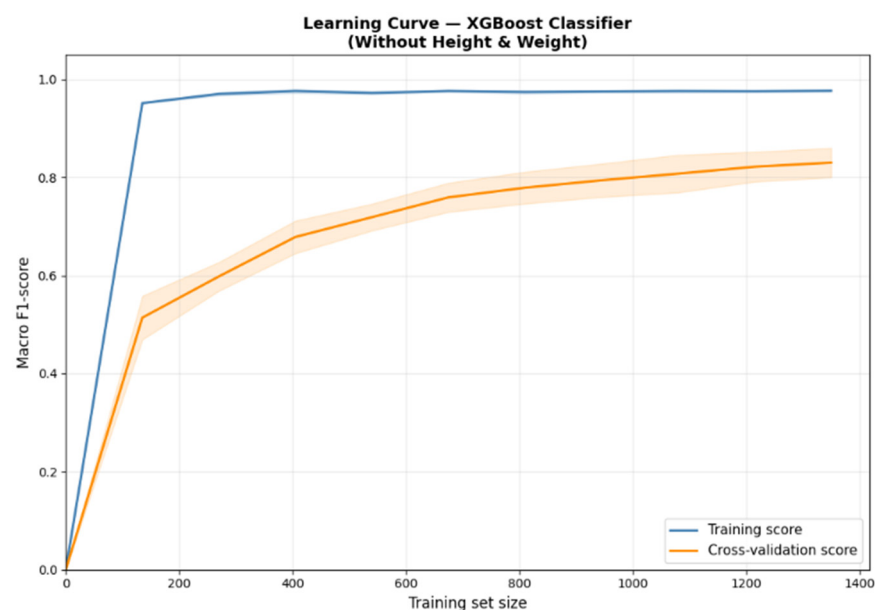


Figure 6. Learning curve showing macro F1-scores on the training and validation sets for increasing training sizes.

To analyze the behavior of the model using more training data, a learning curve was created by plotting the macro F1-score vs. training set size for both the training and validation sets (Figure 6).

The training score initially remains almost constant at about 0.95 and after that, it stabilizes near 0.97 at each training size, showing that XGBoost has a very strong feature to fit the behavioral data over and over again. On the other hand, the cross-validation score

starts at 0.52 when the data are limited and then goes up continuously, eventually reaching almost 0.83 when the full training set has been used. This development of the scores, in step with each other, confirms that the model is substantially benefiting from acquiring new data and at the same time, it is not merely memorizing the training examples.

The difference between two curves becomes smaller step by step along with the increase in training size, which is an indication of a decrease in overfitting and an effective move toward real generalization. The shaded confidence intervals also get tighter with more data, which shows a more stable cross-validation performance. What is more, this generalization is done using only predictors of behavior and lifestyle without any anthropometric ones. In general, the learning curve shows that the XGBoost model is able to generalize well from behavioral data, and it indirectly points out that getting more real-world samples might lead to an even better out-of-sample performance, which is a direction worth exploring further in the nearest future work.

The results presented here tell an important but limited story. They come from a carefully designed classification pipeline evaluated on a single public benchmark dataset, where 77% of the records are synthetically generated. Although the performance is strong, it does not provide evidence that the model would generalize to real clinical populations, and it should not be interpreted in that way. Moving from benchmark performance to clinical usefulness is not automatic; it requires external validation on independent, prospectively collected real-world data from different geographic and demographic contexts. Until such validation is carried out, the main value of this work remains methodological. It offers a rigorous, leakage-free framework for behavioral obesity classification rather than a system ready for clinical deployment.

6.5. Limitations

Some of the limitations of this study should be clearly described.

Firstly, since about 77% of the records in the UCI dataset are synthetically generated, the results presented cannot be assumed to represent real clinical populations accurately. Therefore, the evaluation metrics obtained in this study could be treated as reference results only and not as real-life cases. Also, the synthetic data generator (WEKA) is aimed at simulating the statistical properties and correlational structures of the original data. Although this increases the resemblance to the actual data, it can also lead to strengthening of the behavioral patterns related to obesity in the synthetic dataset. Thus, the residual proxy effects or the indirect encoding of the anthropometric information through behavioral factors, in the absence of explicit body measurements, cannot be completely ruled out.

Secondly, no external validation was done on independent real-world clinical cohorts. The dataset is composed only of the residents from Mexico, Peru, and Colombia, which makes it unsuitable for other geographic areas, cultures, and demographic groups. For this reason, the metrics presented here should not be considered as proof of generalizability outside of this dataset. Also, the validation design does not follow a fully nested cross-validation in the strictest sense of the word. Even though hyperparameter tuning was done within just the training set through an inner five-fold cross-validation loop, to eliminating test set leakage, the use of a single held-out test set tends to give a less stable estimation of the generalization performance compared to a fully nested CV scheme with multiple independent outer folds.

Third, all behavioral variables were self-reported, which might have brought about recall bias and social desirability bias. This can lead to changes in the reliability of the predictors such as eating habits and frequency of exercising.

Fourth, the dataset being cross-sectional means that it is not possible to perform any prospective risk prediction with it. The model that was developed only anticipates the obesity status at present based on the present behaviors; it does not consider a temporal aspect.

Fifth, although SMOTE was used to deal with class imbalance, synthesizing minority sample data may cause noise and weaken the ecological validity of the classification results even more.

Sixth, no assessment of performance in different sex or gender groups was carried out. Obesity-related behavioral factors may vary greatly between males and females, and the model's accuracy might be different in various demographic subgroups. In addition, as gender is one of the predictors, the model may identify gender-related behavioral patterns that probably do not represent the general populations, thus leading to algorithmic bias. In the future, systematic subgroup analyses should be performed to check for fairness and performance equity.

Seventh, this research did not utilize interpretability techniques like SHAP or LIME. Even though the model has very good prediction capabilities, the lack of explanations at the feature level hinders its ready use in clinical decision-making where transparency and interpretability are paramount.

Eighth, issues of data privacy also need to be considered. The UCI dataset, though public and anonymized, brought up concerns about re-identification due to the use of synthetic data generation. Any future use of this framework with actual patient data will require adherence to data protection laws (e.g., GDPR and HIPAA), strong de-identification measures, and formal risk assessment processes.

Ultimately, it is still essential to externally validate the framework on different clinical cohorts and to gather prospective longitudinal data before the real-world implementation of this framework could even be thought of.

6.6. Future Directions

Several avenues have been proposed to make the research more clinically relevant and help it reach the medical setting. One of the most immediate and important actions is to test the method on new, independent groups of patients that are different in location and other characteristics. Confirming the model's effectiveness on actual, soon-to-be-used clinical data, which is not artificially enhanced, is the only way to get a trustworthy indication of the model's generalizability. One more way to bring transparency to the model from a clinical perspective is to use interpretability methods such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations). These methods can show which behavioral features have the most significant impact on making a particular classification at the individual level. Without it, the implementation of clinical decision-support systems might not be possible. Another important step to check the fairness of the model and to see if there is any algorithmic bias towards different population subgroups is to perform subgroup analyses based on sex, gender, age group, and geographic origin. Moreover, by acquiring longitudinal data in a prospective scenario, it will be possible to move from a mere classification of obesity based on present measures to a true prediction of obesity risk, thus pinpointing individuals who are prone to become obese even before changes in their body measurements can be detected. Besides that, a clinical decision-support tool or a web-based calculator, like other ML models in the area of obesity that have been recently suggested, would serve to make the implementation of this system in situations of primary care and public health screening more convenient. Finally, upcoming projects should consider the issues of data privacy and ethical challenges, including fulfilling the data protection rules, choosing the right de-identification methods, and openly communicating the limitations of algorithms to the users at the end of the line.

7. Conclusions

This research created a highly effective machine learning model for classifying obesity status at a given time based on behavioral, demographic, and lifestyle variables. It deliberately excluded anthropometric variables (Height and Weight) to prevent target leakage. The revamped pipeline, which utilized an XGBoost classifier fine-tuned with SMOTE and Optuna hyperparameter tuning (50 trials), obtained a test macro F1-score of 0.872 and a macro AUC of 0.977 on the independent test set, exceeding all baseline classifiers including Random Forest (F1 = 0.869), MLP (F1 = 0.777) and Logistic Regression (F1 = 0.605). These findings indicate that behavioral and lifestyle factors by themselves have a strong discriminative power for classifying obesity status, without reliance on body measurements.

For continued research, this study may be complemented by adding more clinical or socioeconomic features, using explainable AI methods such as SHAP for enhancing interpretability [14], and evaluating sophisticated architectures on larger datasets [15].

External validation on independent clinical data and the integration of multimodal inputs, such as medical notes, would be crucial steps toward developing a more accurate and clinically actionable tool for personalized risk assessment.

Author Contributions: Conceptualization, J.H.; methodology, J.H. and M.E.H.C.; software, J.H.; validation, J.H. and S.A.; formal analysis, S.A.; data curation, J.H.; writing—original draft preparation, J.H. and M.E.H.C.; writing—review and editing, S.A. and M.E.H.C.; visualization, J.H.; project administration, M.E.H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study used a publicly available, anonymized dataset and did not involve human subjects research requiring ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is publicly available at the UCI Machine Learning Repository: <https://archive.ics.uci.edu/> (accessed on 10 February 2026) [13]. The analysis code is available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank the Ibn Tofail University and the Laboratory of Research in Informatics, Kenitra, Morocco, for their institutional support. The authors also extend their sincere gratitude to Salma Azzouzi and Moulay El Hassan Charaf for their valuable guidance and supervision throughout this research, and to Brahim Haggouni for his continued encouragement and support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. World Health Organization. Obesity and Overweight. In *WHO Fact Sheets*; World Health Organization: Geneva, Switzerland, 2021. Available online: <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight> (accessed on 10 February 2026).
2. Blüher, M. Obesity: Global epidemiology and pathogenesis. *Nat. Rev. Endocrinol.* **2019**, *15*, 288–298. [CrossRef] [PubMed]
3. Ferreras, A.; Sumalla-Cano, S.; Martínez-Licort, R.; Elío, I.; Tutusaus, K.; Prola, T.; Vidal-Mazón, J.L.; Sahelices, B.; de la Torre Díez, I. Systematic Review of Machine Learning Applied to the Prediction of Obesity and Overweight. *J. Med. Syst.* **2023**, *47*, 8. [CrossRef] [PubMed]
4. Jamal, H.; Azzimani, K.; Bihri, H.; Salma, A.; El Hassan Charaf, M. Real-Time Cardiovascular Health Monitoring through a Multimodal Data Integration Framework. In Proceedings of the Tenth International Congress on Information and Communication Technology (ICICT 2025), London, UK, 18–21 February 2025. [CrossRef]
5. Rokach, L. Ensemble-based classifiers. *Artif. Intell. Rev.* **2010**, *33*, 1–39. [CrossRef]
6. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]

7. Krawczyk, B. Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* **2021**, *10*, 269–290. [[CrossRef](#)]
8. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
9. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-Generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631. [[CrossRef](#)]
10. Berrar, D. Cross-validation. In *Encyclopedia of Bioinformatics and Computational Biology*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 542–545. [[CrossRef](#)]
11. Jamal, H.; Azzimani, K.; Bihri, H.; Salma, A.; El Hassan Charaf, M. A Comparative Analysis of Random Forest and Decision Tree Classifiers for Predicting Type 2 Diabetes Using K-Fold Cross-Validation. In Proceedings of the 6th International Symposium on Advanced Electrical and Communication Technologies (ISAECT 2024), Alkhobar, Saudi Arabia, 3–5 December 2024; pp. 1–4. [[CrossRef](#)]
12. Probst, P.; Wright, M.N.; Boulesteix, A.-L. Hyperparameters and tuning strategies for random forest. *WIREs Data Min. Knowl. Discov.* **2019**, *9*, e1301. [[CrossRef](#)]
13. UCI Machine Learning Repository. *Obesity Level Estimation Dataset*; University of California: Irvine, CA, USA. Available online: <https://archive.ics.uci.edu/> (accessed on 10 February 2026).
14. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774.
15. Guo, J.; He, Q.; Li, Y. Machine learning-based online web calculator predicts the risk of sarcopenic obesity in older adults. *Aging Clin. Exp. Res.* **2025**, *37*, 210. [[CrossRef](#)] [[PubMed](#)]
16. Ji, C.; Qin, Z.; Yang, Y.; Shen, Y.; Gao, J.; Zhu, F.; Liu, F. Prediction of metabolically healthy obesity based on dietary nutrients: A comparative analysis of six machine learning models with SHAP and LIME interpretation. *Eat. Weight Disord.* **2026**, *31*, 16. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.