

Article

Benchmarking of Ensembles and Meta-Ensembles in the Multiclass Classification of Obesity-Status Classification: Predictive Performance, Calibration and Interpretability

Daniel Andrade-Girón ¹, William Marin-Rodriguez ^{2,*} , Americo Peña ³ , Elsa Oscuvilca-Tapia ⁴ and Fredy Bermejo-Sanchez ³

¹ Department of Formal and Natural Sciences, Universidad Nacional José Faustino Sánchez Carrión, Huacho 15136, Peru; dandrade@unjfsc.edu.pe

² Department of Engineering Systems, Computer and Electronics, Universidad Nacional José Faustino Sánchez Carrión, Huacho 15136, Peru

³ Department of Medicine, Universidad Nacional José Faustino Sánchez Carrión, Huacho 15136, Peru; apena@unjfsc.edu.pe (A.P.); fbermejo@unjfsc.edu.pe (F.B.-S.)

⁴ Department of Nursing, Universidad Nacional José Faustino Sánchez Carrión, Huacho 15136, Peru; eoscuvilca@unjfsc.edu.pe

* Correspondence: wmarin@unjfsc.edu.pe; Tel.: +51-990455214

Abstract

Obesity is a major public health concern because of its high prevalence and association with cardiometabolic comorbidities. This study compared nine ensemble and meta-ensemble learning models for multiclass obesity-status classification using the Obesity Dataset, comprising 1610 records, 14 predictors, and four body-weight status classes. To ensure a leakage-aware evaluation, all preprocessing and resampling steps were embedded within the validation workflow. Standardization, one-hot encoding, and RandomOverSampler were applied only within the training folds; SMOTE and no-resampling configurations were retained as configurable alternatives but were not used to generate the reported results. Model performance was assessed using complementary classification, discrimination, agreement, and calibration metrics, including accuracy, balanced accuracy, weighted F1-score, macro F1-score, weighted ROC-AUC, Matthews correlation coefficient, Brier score, and multiclass expected calibration error. Overall, the ensemble models achieved strong discriminative performance, with eight of nine classifiers exceeding 82% accuracy and obtaining weighted ROC-AUC values close to or above 94%. LightGBM showed the strongest mean metric-based profile, with an accuracy of $85.41 \pm 2.85\%$, weighted F1-score of $85.25 \pm 2.88\%$, weighted ROC-AUC of $95.58 \pm 1.52\%$, and MCC of 0.779 ± 0.042 . Random Forest and Stacking achieved comparable classification performance, although Stacking presented poorer calibration. The Friedman test detected significant global differences among classifiers, $\chi^2 = 38.7733$, $p = 0.000005$. However, the Nemenyi post hoc test indicated that Stacking, Random Forest, LightGBM, Voting, Gradient Boosting, and Extra Trees belonged to the same high-performance statistical group. Therefore, LightGBM was selected as the final model based on its practical balance of predictive performance, calibration behavior, stability, and implementation feasibility, rather than on unequivocal statistical superiority. On the independent holdout set, LightGBM maintained strong generalization, achieving accuracy = 0.8447, weighted F1-score = 0.8435, MCC = 0.7653, and weighted ROC-AUC = 0.9464. Calibration was moderate, with Brier score = 0.2575 and multiclass ECE = 0.1070, indicating that predicted probabilities should be interpreted cautiously when used to support threshold-based decisions.



Academic Editor: Yu Huang

Received: 8 April 2026

Revised: 13 May 2026

Accepted: 1 June 2026

Published: 3 June 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](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: obesity; risk stratification; machine learning; ensemble learning; stacking; Random Forest; Gradient Boosting; explainable AI

1. Introduction

Obesity has emerged as a critical global public health challenge because of its sustained rise, multifactorial etiology and pathogenesis, and close association with chronic noncommunicable diseases [1], including type 2 diabetes mellitus [2], hypertension [3], dyslipidemia, and several types of cancer [4]. Its impact spans the life course, affecting adults, adolescents, and children, and varies across countries and regions according to social, environmental, and behavioral determinants [5]. Consequently, obesity is not only highly prevalent but also a central risk factor for the progression of cardiometabolic comorbidities and systemic complications [6].

Beyond its clinical consequences, obesity imposes a substantial and growing socioeconomic burden. Associated metabolic disorders, including insulin resistance [7], metabolic syndrome, low-grade chronic inflammation, and atherosclerotic cardiovascular disease, increase healthcare utilization, productivity losses, and healthcare expenditures, thereby compromising the sustainability of health systems [8]. Prevalence is projected to continue rising in the coming decades [9]. Recent estimates indicate that more than 1 billion people worldwide are living with obesity, including approximately 650 million adults, 340 million adolescents, and 39 million children [10]. If current trends persist, the annual economic impact of overweight and obesity could reach USD 4.32 trillion by 2035, equivalent to approximately 3% of global GDP [11], and nearly 3.8 billion adults are expected to be obese by 2050 [12]. This scenario underscores the need for cost-effective strategies focused on prevention, early detection, and risk stratification.

In response to this crisis, many countries have implemented population-wide interventions; however, despite their relevance, the benefits have generally been modest and difficult to sustain [13]. In turn, highly effective clinical strategies, such as advanced pharmacotherapy and bariatric surgery, face significant barriers related to cost, access, and scalability, particularly in low- and middle-income countries [14]. Overall, the global response remains fragmented and insufficient relative to the magnitude and pace of the problem [15]. This reinforces the need for comprehensive, multisectoral, and syndemic approaches, supported by analytical tools that can identify at-risk subgroups and prioritize interventions with greater precision.

In this context, machine learning (ML) provides a methodological framework for integrating clinical, biochemical, behavioral, and sociodemographic data to capture nonlinear interactions and generate individualized predictions that support screening and prevention strategies [16]. Recent literature from 2023 to 2025 shows a growing use of predictive applications in cardiometabolic health, particularly models for tabular data and explainability frameworks designed to enhance clinical interpretability [17,18]. However, individual models, such as artificial neural networks (ANNs) [19], support vector machines [20], and decision trees (DTs) [21], have shown important limitations in real-world scenarios characterized by heterogeneity, noise, uncertainty, and class imbalance [22]. Recurrent challenges include overfitting [23], sensitivity to outliers [24], and optimization difficulties, including convergence to local optima in non-convex, high-dimensional spaces, which may reduce performance stability and generalization ability [25].

The challenge is therefore not purely algorithmic. A methodological gap persists in the clinical application of ML to obesity, particularly the lack of standardized frameworks for evaluating the robustness, traceability, and transparency of black-box models trained on

heterogeneous and imbalanced data [26]. This gap may introduce bias when global metrics are prioritized over more informative indicators for multiclass and imbalanced classification, such as the F1-score, area under the curve (AUC)-ROC, and per-class metrics [27]. This problem is further compounded by limited external validation [28], which restricts model transferability to new population contexts. Moreover, the limited integration of interpretability methods, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), into standardized evaluation pipelines hinders the systematic linkage of variable contributions to clinical knowledge, the detection of inconsistencies, the auditing of potential biases, and the assessment of the true utility of these models in public health settings [29].

In light of these limitations, ensemble learning approaches have emerged as effective alternatives for improving predictive stability and generalization [30]. By integrating multiple classifiers through strategies such as bagging, boosting, or stacking [31], ensemble models can reduce variance, mitigate bias, and increase robustness when data are incomplete, heterogeneous, or noisy [32]. Several randomized-tree ensembles also offer relative advantages in interpretability, which is essential for clinical acceptance and potential implementation in public health programs [33]. For contemporary tabular data, gradient-boosting baselines such as XGBoost and LightGBM are necessary competitive benchmarks because they can model nonlinearities and interactions, handle heterogeneity, and maintain strong performance under class imbalance. Therefore, they should be included in methodologically balanced comparisons [34,35].

Within this framework, the present study aims to develop, evaluate, and compare nine ensemble and meta-ensemble models for multiclass obesity-status classification. Model performance was assessed using a two-stage evaluation design consisting of an independent stratified holdout set for final generalization testing and nested stratified cross-validation within the development set. Specifically, the development data were evaluated using a 10-fold outer loop for performance estimation and a 3-fold inner loop for hyperparameter optimization. Performance was quantified using complementary classification, discrimination, agreement, and calibration metrics, including accuracy, balanced accuracy, weighted and macro F1-scores, weighted ROC-AUC, Matthews correlation coefficient, Brier score, and multiclass expected calibration error.

To strengthen the comparison, the analytical framework will incorporate non-parametric statistical tests, including the Friedman test and Nemenyi post hoc analysis, to detect significant differences among models. In addition, global interpretability using SHAP and local explainability using LIME will be integrated to assess clinical coherence, identify relevant predictors, explain individual-level decisions, and explore potential patterns of bias.

Accordingly, this study proposes a predictive, explainable, and interpretable architecture organized around three main contributions. First, it presents a reproducible, leakage-aware methodological protocol for multiclass obesity-status classification, integrating preprocessing, class-imbalance handling, hyperparameter optimization, nested stratified cross-validation, and independent holdout testing. Second, it compares nine ensemble learning methods—LightGBM, Random Forest, Stacking, Voting, Gradient Boosting, Extra Trees, Bagging, XGBoost, and AdaBoost—that span bagging, boosting, and meta-ensemble strategies, using a comprehensive set of performance and agreement metrics, including the Matthews correlation coefficient (MCC). Third, it incorporates a global-local interpretability framework based on SHAP and LIME, aimed at auditing clinical consistency, identifying relevant population-level predictors, explaining individual predictions, and facilitating model transfer to applied screening and clinical decision-support scenarios.

2. Literature Review

Obesity remains a major challenge for contemporary public health because of its multifactorial and multisystemic etiology. In this context, machine learning (ML) offers a robust alternative for early risk prediction, particularly when nonlinear relationships, high-dimensional interactions, collinearity, noise, and class imbalance limit the performance of conventional statistical approaches. By modeling complex patterns across heterogeneous data sources, ML can support clinically meaningful risk stratification and early preventive decision-making.

Dugan et al. [36] provided one of the earliest examples of predicting childhood obesity before age two using clinical cohorts and classical algorithms, including ID3, Random Forest, and Naive Bayes. Although innovative, the study relied primarily on conventional performance measures and did not include comparative statistical testing, thereby limiting the inferential strength of its findings.

Since 2020, the literature has shown a marked increase in methodological complexity, with greater emphasis on interpretable ML, electronic health records (EHRs), and large-scale clinical data. Fu et al. [37] incorporated LightGBM and SHAP analysis in a cohort of preterm infants, identifying early body mass index (BMI) trajectories and maternal BMI as relevant predictive factors. Complementarily, Colmenarejo [2] examined the use of ML and deep learning in pediatric EHRs, reporting superior performance relative to classical statistical regression while also emphasizing persistent deficiencies in clinical validation.

Other studies expanded both the scale and methodological diversity of obesity prediction. Ferdowsy et al. [3] reported that Logistic Regression achieved the highest accuracy (97.09%) among the evaluated classifiers, whereas Gradient Boosting showed the lowest performance, with an accuracy of 64.08% and lower values across the remaining metrics. Shi et al. [38] applied a Bagging-based Feature Selection framework integrating MapReduce (BFSMR) and weighted feature voting to identify informative clinical factors among more than 426,000 patients. Pang et al. [39] used XGBoost on more than 800,000 pediatric visits and achieved an AUC of 0.81 for future obesity prediction. Safaei et al. [7], through a systematic review of 93 studies, further confirmed the heterogeneity of algorithms, datasets, and validation strategies in the field. Despite their scale and relevance, these studies generally did not incorporate adjusted *p*-values or multiple statistical comparisons.

By 2022, research had shifted toward more advanced architectures. Cheng et al. [40] employed long short-term memory (LSTM) networks to predict BMI in pediatric cohorts with missing values, whereas Liu et al. [41] used deep reinforcement learning to identify gut microbiome biomarkers associated with obesity. These contributions illustrate the expansion of obesity prediction beyond classical classifiers toward sequential, biological, and reinforcement learning frameworks.

Recent studies from 2023 onward provide further evidence for the predictive value of ensemble and hybrid models. Solomon et al. [42] proposed a hybrid majority-voting model combining Gradient Boosting, XGBoost phylogenetic tree algorithms, and multi-layer perceptron (MLP) artificial neural networks, achieving 97.16% accuracy on intensive care unit (ICU) datasets. Abnoosian et al. [43] evaluated multiple classifiers in patients with diabetes and reported an AUC close to 1.0, a promising result that nevertheless underscores the need for external validation and bias assessment. Matt et al. [44] compared 10 algorithms and identified Random Forest and Logistic Regression as the most robust models. Overall, these studies report high predictive performance, but their validity remains constrained by limited multicenter external validation and insufficient statistical testing to determine whether differences between algorithms are significant.

More recent work has increasingly focused on hybrid modeling and interpretability. Helforouh and Sayyad [45] proposed an ANN-particle swarm optimization (PSO)

approach enhanced with SHAP, reporting improvements over conventional regression methods. Sun et al. [46] applied Gradient Boosted Decision Trees (GBDT) to large international cohorts, including the China Health and Nutrition Survey (CHNS) and the National Health and Nutrition Examination Survey (NHANES), identifying physical inactivity and alcohol consumption as key determinants in obesity prediction. Airlangga [17] evaluated eight ML algorithms using self-reported data from Mexico, Peru, and Colombia, showing that Gradient Boosting achieved the best performance (96.49%) and effectively captured complex demographic, behavioral, and lifestyle interactions, while simpler classifiers such as Naive Bayes and AdaBoost performed less favorably. In addition, Chatterjee et al. [47] developed a hybrid ensemble Stacking model that achieved 98.58% accuracy, with consistent F1-score, AUC, and kappa values (kappa = 0.9834), demonstrating superior performance over individual classifiers and potential applicability in public health decision-support systems.

Overall, the reviewed literature shows a clear shift from classical algorithms toward more complex, hybrid, and interpretable architectures, generally accompanied by improvements in predictive performance, robustness, and generalizability. However, this evolution remains constrained by methodological heterogeneity that limits comparability and inferential scope. Adjusted agreement metrics, particularly the Matthews correlation coefficient (MCC), remain underreported despite their relevance in imbalanced or multiclass classification settings. Likewise, nonparametric procedures such as the Friedman test with Nemenyi post hoc comparisons are rarely used to assess whether observed differences among algorithms, including contemporary Gradient Boosting-based baselines, are statistically significant under a comparable evaluation protocol.

Accordingly, a comprehensive benchmarking framework is needed to rigorously, reproducibly, and comparably assess the performance of machine learning models for the multiclass classification of obesity-status classification. Such a framework should include competitive ensemble models, such as XGBoost, LightGBM, and other approaches based on bagging, boosting, and meta-ensemble strategies; complementary performance and agreement metrics, including the Matthews correlation coefficient (MCC); non-parametric statistical tests, such as the Friedman and Nemenyi tests, to compare differences among models; and a global–local interpretability scheme based on SHAP and LIME. Such an integrated approach would enable comparable, and clinically interpretable conclusions, thereby strengthening the utility of these models in screening and clinical decision-support scenarios related to obesity-status classification.

3. Methodology

3.1. Study Design and Data Source

The dataset analyzed in this study was obtained from a publicly available Kaggle resource on obesity-status classification: <https://www.kaggle.com/code/zachorwin/data-science-for-change#Introduction-and-Problem-Statement>. The resource was accessed on 27 June 2025. The analytical file used in this study contained 1610 observations, 14 predictors, and one multiclass target variable. No new data were generated. The variable structure and data-collection context are consistent with the obesity dataset described by Köklü and Sulak in the Sinop University Journal of Natural Sciences (2024) [48].

3.2. Outcome Definition and Predictor Variables

The analytical dataset comprised 15 variables: 14 predictors and one multiclass target variable. The predictors included demographic, anthropometric, family history, dietary, behavioral, and transportation-related variables: Sex, Age, Height, Overweight_Obese_Family, Consumption_of_Fast_Food, Frequency_of_Consuming_Vegetables,

Number_of_Main_Meals_Daily, Food_Intake_Between_Meals, Smoking, Liquid_Intake_Daily, Calculation_of_Calorie_Intake, Physical_Exercise, Time_Spent_on_Technology, and Type_of_Transportation_Used. The target variable, Class, represented body-weight status with four ordered categories: Underweight = 1, Normal = 2, Overweight = 3, and Obesity = 4. The class distribution was 73 underweight cases (4.5%), 658 normal-weight cases (40.9%), 592 overweight cases (36.8%), and 287 obesity cases (17.8%). This distribution motivated the use of stratified data partitioning, imbalance-aware resampling, and performance metrics that are more informative than overall accuracy alone.

3.3. Reproducible Computational Environment

All experiments were conducted in Python version 3.10.12 (Python Software Foundation, Wilmington, DE, USA) within a cloud-based computational environment using Google Colaboratory, also known as Google Colab (Google LLC, Mountain View, CA, USA). The analyses were implemented using a reproducible pipeline architecture designed to ensure consistency across preprocessing, model training, hyperparameter optimization, and evaluation.

To support reproducibility and ensure a fair comparison across models, the same Google Colab environment configuration and a fixed random seed (`random_state = 42`) were used across all experiments. Computational runtime, measured in seconds, included model fitting, hyperparameter search, and evaluation within the implemented pipeline.

3.4. Data Cleaning and Preprocessing

Column names were standardized before analysis by removing leading/trailing spaces and replacing spaces, hyphens, slashes, parentheses, percentage signs, and periods with machine-readable tokens. Duplicate records were removed using `drop_duplicates()`, and completely empty columns, constant predictors, and identifier-like columns with very high cardinality were excluded to reduce the risk of spurious predictive signals. The target column was validated explicitly and encoded with `LabelEncoder`. Categorical predictors were preserved as categorical variables, whereas explicitly defined continuous or ordinal predictors were coerced to numeric format. Although the exploratory inspection did not indicate substantive missingness in the target, preprocessing transformers were retained inside the modelling pipeline to guarantee robustness, and reproducibility should missing predictor values be encountered during resampling or cross-validation.

3.5. Leakage-Controlled Preprocessing Pipeline

To prevent information leakage, all transformations were embedded within a single imbalanced-learn pipeline and fitted exclusively within the training partition of each validation fold. Numerical predictors were processed using a median imputer followed by `StandardScaler`. Categorical predictors were processed using a most-frequent imputer followed by `OneHotEncoder` with `handle_unknown = 'ignore'`. These transformations were implemented through a `ColumnTransformer` and applied identically to all candidate models. Embedding preprocessing within the cross-validation pipeline ensured that imputation, scaling, one-hot encoding, resampling, and model fitting were estimated only from the training data within each fold and then applied to the corresponding validation fold.

3.6. Handling of Class Imbalance

Class imbalance was addressed exclusively within the training folds to prevent information leakage and to preserve the validity of out-of-sample estimates. In the reported experimental configuration, `RandomOverSampler` was used as the resampling strategy, whereas `SMOTE` and no-resampling settings were retained only as configurable alternatives and were not used to generate the results reported in this study. This clarification ensures

consistency between the methodological description, the experimental pipeline, and the reported performance estimates.

Standardization, one-hot encoding, and resampling were embedded within a single imbalanced-learn pipeline, ensuring that each preprocessing and balancing step was fitted only on the training subset of each cross-validation fold. Consequently, replicated observations produced by RandomOverSampler were generated exclusively from the training data and were never introduced into the validation folds or the independent holdout set. This leakage-controlled design preserved the independence of validation and holdout samples, reduced the risk of selection-induced optimism, and strengthened the reproducibility and methodological transparency of the evaluation workflow.

3.7. Candidate Learning Algorithms

A broad set of ensemble and gradient-boosting algorithms was evaluated to capture nonlinear relationships, high-order interactions, and heterogeneous predictor effects. The candidate models included AdaBoost, Gradient Boosting, Random Forest, Extra Trees, Bagging, soft Voting, Stacking, XGBoost and LightGBM. AdaBoost was included as a boosting-based ensemble that sequentially reweights misclassified observations and combines weak learners into a stronger classifier [49–51]. Gradient Boosting was evaluated because it builds additive models by sequentially minimizing a differentiable loss function and correcting residual errors from previous learners [50]. Random Forest was included as a bootstrap-aggregation method based on randomized decision trees, which can reduce variance and mitigate overfitting [52,53]. Extra Trees introduced additional randomization at the split-threshold level to increase tree diversity and improve robustness to noise [54]. Bagging was evaluated as a variance-reduction strategy based on bootstrap sampling and majority voting [55].

3.8. Hyperparameter Optimization

Hyperparameter tuning was performed within the inner loop of the nested validation procedure. For each model, the search space was defined a priori and passed to either GridSearchCV or RandomizedSearchCV. Small, bounded spaces were evaluated with grid search, whereas larger spaces were evaluated with randomized search using up to 50 candidate configurations. The primary tuning metric was F1-score for binary tasks and weighted F1-score for multiclass tasks. Weighted F1 was selected for the present multiclass obesity setting because it accounts for class support while balancing precision and recall. The search was refitted on the corresponding training fold using the best-performing hyperparameter configuration before evaluation on the held-out outer fold.

3.9. Evaluation Design: Nested Cross-Validation and Independent Holdout Testing

A two-stage evaluation design was implemented to separate model selection from final generalization assessment. First, the complete dataset was split via stratified sampling into a development set comprising 80% of observations and an independent holdout set comprising 20%. The holdout set was kept isolated from all preprocessing decisions, resampling, hyperparameter tuning, model selection, and calibration. Second, all model development and benchmarking procedures were conducted only within the development set using nested stratified cross-validation. The outer loop used 10 folds to obtain an approximately unbiased estimate of model performance, whereas the inner loop used 3 folds for hyperparameter optimization.

3.10. Performance Metrics

Model performance was quantified using complementary discrimination, agreement, and probabilistic calibration metrics. Discrimination metrics included accuracy, balanced

accuracy, and weighted and macro F1-score, precision, and recall, as well as weighted and macro one-vs-rest ROC-AUC when applicable. Agreement was assessed using the Matthews correlation coefficient (MCC), which is particularly useful when class distributions are imbalanced or when multiclass errors must be summarized beyond accuracy. Probabilistic performance was evaluated with the Brier score and expected calibration error (ECE). For multiclass classification, ROC-AUC was computed using a one-vs-rest strategy with macro and weighted averaging, and multiclass ECE was based on the maximum predicted class confidence. Cross-validation results were summarized as mean $\pm$ standard deviation across outer folds, whereas holdout results were reported as single independent estimates for the final selected model.

3.11. Statistical Comparison of Algorithms

To avoid interpreting differences among algorithms solely from average performance, nonparametric statistical comparisons were conducted across the outer cross-validation folds. The Friedman test was applied to the weighted F1-score distributions to assess whether statistically significant differences existed among the candidate models evaluated under the same fold partitions. When post hoc analysis was available, Nemenyi pairwise comparisons were computed to identify which model pairs differed significantly while controlling the family-wise error rate. Average ranks were reported to summarize the relative ordering of models across folds, and compact letter displays were generated to facilitate interpretation of statistically indistinguishable model groups.

3.12. Model Interpretability and Explainability

Model interpretability was assessed after model development and final fitting, rather than during hyperparameter tuning. SHAP analysis was implemented using class-specific one-vs-rest binary LightGBM models, and SHAP values were computed on the log-odds scale for real observations from the independent test/holdout partition, after transformation by the fitted preprocessing workflow. LIME was used as a complementary local explanation method for the multiclass LightGBM model. The analyzed LIME instance was selected from the test/holdout partition using the 85th confidence quantile of the calibrated class probabilities. The probability panel reported post hoc calibrated probabilities, whereas the LIME contribution plot approximated the local behavior of the underlying multiclass LightGBM classifier.

Oversampling was applied only during model training and not to the observations analyzed with SHAP or LIME. Thus, the interpretability analyses were performed on real, non-synthetic test/holdout observations. Nevertheless, because the models were trained on a rebalanced training distribution and the LIME explainer used the upsampled training representation, the resulting attribution patterns may still have been indirectly influenced by the upsampling procedure. SHAP and LIME outputs were therefore interpreted as model-dependent predictive associations rather than causal effects or clinically validated decision rules.

4. Results

The proposed ensemble and meta-ensemble benchmarking framework enabled a structured evaluation of multiclass obesity-status classification under a unified, reproducible, and leakage-controlled experimental protocol. The analysis was conducted on a heterogeneous clinical-epidemiological feature space composed of demographic, anthropometric, familial, dietary, lifestyle, and transportation-related predictors. Given the expected presence of nonlinear associations, high-order interactions, and class-imbalanced distributions in obesity-related data, the evaluation emphasized ensemble-based strategies with

the potential to improve discrimination stability, variance control, and robustness when compared with single-estimator approaches. Model performance was assessed within a nested stratified cross-validation framework, and the independent holdout set was reserved exclusively for the final assessment of generalization performance.

The data integrity procedure confirmed that the analytical workflow was applied consistently across all predictors, including Sex, Age, Height, Overweight_Obese_Family, Consumption_of_Fast_Food, Frequency_of_Consuming_Vegetables, Number_of_Main_Meals_Daily, Food_Intake_Between_Meals, Smoking, Liquid_Intake_Daily, Calculation_of_Calorie_Intake, Physical_Exercise, Time_Spent_on_Technology, and Type_of_Transportation_Used. Prior to model training, column names were standardized, duplicate observations were removed, empty variables were excluded when detected, and identifier-like variables were screened to reduce the risk of spurious memorization. The target variable, Class, was encoded as a multiclass outcome. Although the exploratory missingness assessment indicated complete data, imputation procedures were retained within the preprocessing pipeline as a reproducibility safeguard. Numerical predictors were processed using median imputation followed by standardization, whereas categorical predictors were processed using most-frequent imputation followed by one-hot encoding with unknown categories ignored during transformation. This design ensured that the same preprocessing logic was applied within each cross-validation split and during final holdout evaluation.

To preserve the validity of out-of-sample estimates, preprocessing, resampling, and classification were encapsulated in a single imblearn pipeline. Class imbalance was addressed with RandomOverSampler in its default configuration, and the resampling step was fitted only on the training portion of each cross-validation split. This prevented information from validation or holdout observations from influencing imputation, scaling, encoding, resampling, hyperparameter tuning, or model fitting. As a result, the benchmarking process provided a leakage-controlled comparison of the candidate models and reduced the risk of overly optimistic performance estimates.

The evaluated algorithms were grouped by their ensemble learning strategy. Bagging-based methods included Random Forest, Extra Trees, and Bagging with decision trees, while boosting-based methods included AdaBoost, Gradient Boosting, XGBoost, and LightGBM, when the corresponding libraries were available. Meta-ensemble approaches included Stacking, implemented with a regularized Logistic Regression meta-learner, and soft Voting, based on probabilistic aggregation of predictions from Random Forest, Extra Trees, and Gradient Boosting. This classification provides a more accurate methodological interpretation of the evaluated algorithms, particularly because gradient-boosted tree methods are themselves ensemble models and should not be described as non-ensemble baselines.

The predictive performance of the models was evaluated using complementary metrics capturing discrimination, class-balance sensitivity, agreement, and probabilistic reliability. Accuracy was considered together with balanced accuracy, macro and weighted F1-score, macro and weighted precision and recall, macro and weighted ROC-AUC, Matthews correlation coefficient, Brier score, and expected calibration error. This metric set allowed the analysis to move beyond overall accuracy and to account for multiclass imbalance, agreement beyond chance, and the reliability of predicted probabilities. Unless otherwise specified, cross-validation results should be interpreted as mean $\pm$ standard deviation across the outer folds.

Hyperparameter optimization was performed within the inner loop of the nested cross-validation design. The outer loop used ten-fold stratified cross-validation for unbiased performance estimation, whereas the inner loop used three-fold stratified cross-validation for model-specific tuning. GridSearchCV was used for small, bounded search spaces, whereas RandomizedSearchCV was used for larger spaces, evaluating up to fifty config-

urations when appropriate. For the multiclass task, the weighted F1-score was used as the primary optimization criterion. Final model selection followed a predefined hierarchical rule, first based on the mean weighted F1-score, then on the Matthews correlation coefficient, and finally on the weighted ROC-AUC.

After model selection, the final estimator was refitted on the development data and evaluated once on the independent holdout set. When probabilistic calibration was enabled, the selected model was calibrated using sigmoid calibration with three-fold cross-validation on the development data. Calibration performance was assessed using the multiclass Brier score, expected calibration error, global reliability diagrams based on maximum predicted confidence, and class-wise reliability diagrams. These analyses provided evidence on whether the predicted probabilities were numerically consistent with observed outcomes, which is particularly relevant for clinical and public-health decision-support scenarios.

Model interpretability for tree-based estimators was assessed using SHAP values computed on a predefined sample of observations. Mean absolute SHAP values were used to rank the most influential transformed features, thereby complementing the discrimination and calibration results with evidence on model transparency. This interpretability analysis allowed the benchmarking framework to assess whether the most relevant predictors were clinically plausible and consistent with obesity-related demographic, anthropometric, familial, dietary, and lifestyle factors. Overall, the framework integrates predictive performance, robustness, probabilistic reliability, and interpretability, providing a rigorous basis for evaluating ensemble and meta-ensemble models in multiclass obesity-status classification.

As illustrated in Table 1, the results obtained for each model are accompanied by a critical analysis that examines their relative performance in terms of accuracy, robustness, and generalizability. This analysis provides a solid foundation for determining optimal configurations in practical applications and future lines of research.

Table 1. Main results by model in nested cross-validation. For Brier score and ECE, lower values indicate better calibration.

Modelo	Accuracy (%)	BalAcc (%)	F1 Weighted (%)	F1 Macro (%)	AUC Weighted (%)	MCC	Brier	ECE mc
LightGBM	85.41 ± 2.85	81.29 ± 2.69	85.25 ± 2.88	83.30 ± 2.58	95.58 ± 1.52	0.779 ± 0.042	0.2451 ± 0.0469	0.1083 ± 0.0281
RandomForest	85.17 ± 2.02	83.23 ± 3.71	85.07 ± 2.01	83.99 ± 3.50	95.29 ± 1.52	0.777 ± 0.031	0.2495 ± 0.0216	0.1242 ± 0.0208
Stacking	85.10 ± 2.79	81.51 ± 4.23	84.97 ± 2.84	83.36 ± 3.91	95.12 ± 1.49	0.775 ± 0.042	0.4061 ± 0.1586	0.2965 ± 0.2100
Voting	84.63 ± 3.04	81.96 ± 3.23	84.54 ± 3.05	82.91 ± 3.44	95.11 ± 1.52	0.769 ± 0.045	0.2480 ± 0.0262	0.0971 ± 0.0265
GradientBoosting	84.32 ± 4.08	81.88 ± 4.36	84.26 ± 4.10	82.69 ± 4.34	94.75 ± 1.65	0.764 ± 0.061	0.2603 ± 0.0618	0.1115 ± 0.0344
ExtraTrees	83.85 ± 3.18	79.69 ± 5.37	83.70 ± 3.18	81.21 ± 5.27	94.83 ± 1.52	0.757 ± 0.047	0.2489 ± 0.0266	0.0842 ± 0.0189
Bagging	83.78 ± 2.65	80.56 ± 4.12	83.62 ± 2.67	81.84 ± 3.80	94.94 ± 1.91	0.756 ± 0.039	0.2681 ± 0.0250	0.1370 ± 0.0292
XGBoost	82.77 ± 2.91	80.42 ± 3.42	82.76 ± 2.82	80.15 ± 3.96	94.78 ± 1.59	0.742 ± 0.042	0.2532 ± 0.0371	0.0748 ± 0.0118
AdaBoost	66.46 ± 3.12	67.83 ± 5.07	67.12 ± 3.16	61.02 ± 3.52	85.88 ± 1.37	0.519 ± 0.041	0.7291 ± 0.0025	0.4022 ± 0.0312

Note. Table 1 summarizes the predictive performance of the evaluated ensemble and meta-ensemble learning models under the nested cross-validation scheme. Overall, most classifiers showed strong and stable discriminative performance, with accuracy values above 82% for eight of the nine models and weighted ROC-AUC values close to or above 94%. These findings indicate that ensemble-based methods provided robust discriminative capacity for the multiclass obesity-status classification task.

Among the evaluated algorithms, LightGBM achieved the strongest overall metric-based profile, obtaining the highest accuracy of 85.41 ± 2.85%, the highest weighted F1-score of 85.25 ± 2.88%, a macro F1-score of 83.30 ± 2.58%, the highest weighted ROC-AUC of 95.58 ± 1.52%, and the highest MCC value of 0.779 ± 0.042. In addition, LightGBM obtained the lowest Brier score among the evaluated models, 0.2451 ± 0.0469, suggesting a favorable balance between discrimination, agreement, and probabilistic performance. These results support LightGBM as the model with the most favorable overall metric-based profile under the proposed nested cross-validation framework.

Random Forest ranked second and showed the strongest class-balanced behavior. It achieved an accuracy of 85.17 ± 2.02%, the highest balanced accuracy among all models, 83.23 ± 3.71%, the highest macro F1-score, 83.99 ± 3.50%, a weighted F1-score of

85.07 $\pm$ 2.01%, and an MCC of 0.777 $\pm$ 0.031. Although its global accuracy was slightly lower than that of LightGBM, Random Forest demonstrated a particularly robust ability to maintain performance across the four obesity-status classes. This is relevant in multiclass settings, where balanced class-level performance is preferable to high accuracy driven mainly by majority classes.

The Stacking ensemble also achieved highly competitive discrimination metrics, with an accuracy of 85.10 $\pm$ 2.79%, a weighted F1-score of 84.97 $\pm$ 2.84%, a macro F1-score of 83.36 $\pm$ 3.91%, a weighted ROC-AUC of 95.12 $\pm$ 1.49%, and an MCC of 0.775 $\pm$ 0.042. However, its calibration metrics were notably weaker, with a Brier score of 0.4061 $\pm$ 0.1586 and a multiclass ECE of 0.2965 $\pm$ 0.2100. This indicates that, despite its strong discriminative performance, the probability estimates produced by Stacking were less reliable than those of other high-performing models. Therefore, its use in threshold-based clinical or operational decisions would require additional calibration assessment.

The Voting ensemble ranked fourth and showed a favorable compromise between predictive performance and calibration. It achieved an accuracy of 84.63 $\pm$ 3.04%, a balanced accuracy of 81.96 $\pm$ 3.23%, a weighted F1-score of 84.54 $\pm$ 3.05%, a weighted ROC-AUC of 95.11 $\pm$ 1.52%, and an MCC of 0.769 $\pm$ 0.045. Its calibration profile was also favorable, with a Brier score of 0.2480 $\pm$ 0.0262 and a multiclass ECE of 0.0971 $\pm$ 0.0265. However, Voting did not obtain the lowest Brier score, since LightGBM showed a slightly lower value. Thus, Voting should be interpreted as one of the best-calibrated alternatives, but not as the best model according to the Brier score criterion.

Gradient Boosting ranked fifth, with an accuracy of 84.32 $\pm$ 4.08%, a balanced accuracy of 81.88 $\pm$ 4.36%, a weighted F1-score of 84.26 $\pm$ 4.10%, a weighted ROC-AUC of 94.75 $\pm$ 1.65%, and an MCC of 0.764 $\pm$ 0.061. Although its discriminative performance remained high, the larger standard deviations suggest greater variability across validation folds compared with LightGBM and Random Forest. This indicates that Gradient Boosting was competitive but somewhat less stable under the evaluated resampling partitions.

Extra Trees and Bagging showed competitive but slightly lower performance than the leading models. Extra Trees achieved an accuracy of 83.85 $\pm$ 3.18%, a weighted F1-score of 83.70 $\pm$ 3.18%, a macro F1-score of 81.21 $\pm$ 5.27%, a weighted ROC-AUC of 94.83 $\pm$ 1.52%, and an MCC of 0.757 $\pm$ 0.047. Bagging obtained an accuracy of 83.78 $\pm$ 2.65%, a weighted F1-score of 83.62 $\pm$ 2.67%, a macro F1-score of 81.84 $\pm$ 3.80%, a weighted ROC-AUC of 94.94 $\pm$ 1.91%, and an MCC of 0.756 $\pm$ 0.039. Although neither model reached the top-ranked performance levels, both maintained weighted ROC-AUC values close to 95%, confirming their capacity to distinguish among obesity-status categories with reasonable reliability.

XGBoost ranked eighth in terms of global predictive performance, with an accuracy of 82.77 $\pm$ 2.91%, a balanced accuracy of 80.42 $\pm$ 3.42%, a weighted F1-score of 82.76 $\pm$ 2.82%, a macro F1-score of 80.15 $\pm$ 3.96%, a weighted ROC-AUC of 94.78 $\pm$ 1.59%, and an MCC of 0.742 $\pm$ 0.042. Despite its lower position in the global performance ranking, XGBoost achieved the lowest multiclass ECE among all models, 0.0748 $\pm$ 0.0118. This suggests that XGBoost provided the most favorable calibration according to expected calibration error, although this advantage did not translate into the highest discrimination or agreement metrics.

In contrast, AdaBoost showed the weakest performance across all evaluated metrics. It obtained an accuracy of 66.46 $\pm$ 3.12%, a balanced accuracy of 67.83 $\pm$ 5.07%, a weighted F1-score of 67.12 $\pm$ 3.16%, a macro F1-score of 61.02 $\pm$ 3.52%, and the lowest MCC value, 0.519 $\pm$ 0.041. It also presented the poorest calibration, with the highest Brier score, 0.7291 $\pm$ 0.0025, and the highest multiclass ECE, 0.4022 $\pm$ 0.0312. These results suggest that AdaBoost was less effective in capturing the nonlinear and multiclass structure of the obesity-status classification problem.

Overall, LightGBM achieved the strongest mean metric-based profile, with the highest accuracy, weighted F1-score, weighted ROC-AUC, MCC, and the lowest Brier score. Random Forest showed the strongest class-balanced behavior, whereas XGBoost obtained the lowest multiclass ECE. Voting also displayed favorable calibration, although it did not achieve the best Brier score or the lowest ECE. These results indicate that model preference depends on the intended use: LightGBM may be selected when overall predictive performance and balanced probabilistic behavior are prioritized; Random Forest when balanced class-level performance is required; XGBoost when minimizing multiclass ECE is the main priority; and voting when a balanced compromise between discrimination and calibration is desired. However, descriptive differences in mean metrics are insufficient to establish statistical superiority; therefore, Friedman and Nemenyi post hoc analyses were conducted to formally compare the classifiers, as summarized in Table 2.

Table 2. Friedman ranking and Nemenyi post hoc grouping of ensemble models based on weighted F1-score.

Position	Model	Mean F1w	SD	Average Rank	Nemenyi Group
1	Stacking	0.850	0.028	3.450	A
2	RandomForest	0.851	0.020	3.500	A
3	LightGBM	0.853	0.029	3.800	A
4	Voting	0.845	0.031	3.950	A
5	GradientBoosting	0.843	0.041	4.100	A
6	ExtraTrees	0.837	0.032	5.000	A
7	Bagging	0.836	0.027	5.800	A-B
8	XGBoost	0.828	0.028	6.400	A-B
9	AdaBoost	0.671	0.032	9.000	B

Note. Mean F1w denotes the mean weighted F1-score across the outer cross-validation folds. Average ranks were derived from the Friedman test, and statistical groups were obtained using the Nemenyi post hoc comparison. Lower average ranks indicate better relative performance across folds. Models sharing at least one letter in the Nemenyi group column are not significantly different under the post hoc comparison.

The Friedman test showed statistically significant global differences among the evaluated classifiers in terms of the weighted F1-score, $\chi^2 = 38.7733$, $p = 0.000005$. This result indicates that the classifiers did not perform equivalently across the cross-validation folds and justifies using a post hoc Nemenyi analysis to identify groups of statistically comparable models.

The Nemenyi ranking analysis showed that Stacking achieved the best average rank among the evaluated methods, with an average rank of 3.450, followed closely by RandomForest (3.500), LightGBM (3.800), Voting (3.950), and GradientBoosting (4.100). These models were assigned to the same Nemenyi group, group A, indicating that their rank differences were not statistically significant under the post hoc comparison. ExtraTrees also belonged to group A, with an average rank of 5.000, suggesting that its fold-wise performance was not significantly different from that of the leading classifiers.

Although Stacking achieved the most favorable average rank, LightGBM obtained the highest mean weighted F1-score (0.853 ± 0.029), followed by RandomForest (0.851 ± 0.020) and Stacking (0.850 ± 0.028). The differences among these top-performing models were small, supporting the presence of a compact high-performance cluster rather than a single classifier with unequivocal statistical dominance. Therefore, the apparent superiority of LightGBM should be interpreted as a practical advantage in terms of mean weighted F1-score, rather than as statistically significant dominance over the other high-ranking models.

Voting and GradientBoosting also showed competitive performance, with mean weighted F1-scores of 0.845 ± 0.031 and 0.843 ± 0.041 , respectively. Both models remained within Nemenyi group A, reinforcing the conclusion that their predictive performance

was statistically comparable to that of the leading algorithms. ExtraTrees achieved a mean weighted F1-score of 0.837 ± 0.032 and was also included in group A, although its average rank was less favorable than those of Stacking, RandomForest, LightGBM, Voting, and GradientBoosting.

Bagging and XGBoost occupied intermediate positions. Bagging obtained a mean weighted F1-score of 0.836 ± 0.027 and an average rank of 5.800, whereas XGBoost achieved 0.828 ± 0.028 with an average rank of 6.400. Both models were assigned to the overlapping A–B group, indicating that their performance was not clearly separated from either the higher-performing group or the lower-performing group. This overlapping classification suggests that, although their average ranks were less favorable, there was insufficient evidence to classify them as significantly worse than all group A models.

By contrast, AdaBoost showed the weakest performance, with the lowest mean weighted F1-score, 0.671 ± 0.032 , and the worst average rank, 9.000. It was the only model assigned exclusively to group B, confirming its consistently poorer relative performance across folds. This result indicates that AdaBoost was statistically separated from the strongest-performing classifiers and was less suitable for the multiclass prediction task considered in this study.

Overall, the non-parametric analysis supports the presence of significant differences among classifiers, while the Nemenyi grouping indicates that the best-performing models formed a statistically comparable high-performance cluster. In this context, LightGBM was selected as the final model on practical grounds, as it achieved the highest mean weighted F1-score while maintaining a competitive average rank and belonging to the top Nemenyi group. However, this selection should not be interpreted as evidence of statistically significant superiority over Stacking, RandomForest, Voting, GradientBoosting, or ExtraTrees. Rather, LightGBM was preferred because it provided the best balance between mean predictive performance, ranking stability, and practical model selection criteria.

The final LightGBM model was subsequently trained on the nested-training set and evaluated once on the independent holdout set. This strategy preserved the independence of the test set and reduced the risk of selection-induced optimism. In the holdout evaluation, the model maintained strong predictive performance, with an accuracy of 0.8447, a weighted F1-score of 0.8435, a macro F1-score of 0.8022, an MCC of 0.7653, and a weighted ROC-AUC of 0.9464. These results indicate that the selected model retained good discrimination and generalization ability when assessed on previously unseen data.

The final model calibration yielded a multiclass Brier score of 0.2575 and a multiclass ECE of 0.1070, suggesting a moderate overall calibration performance. The reliability diagram showed that the predicted probabilities were reasonably consistent with the observed accuracy in the highest-confidence interval (0.8–0.9). However, discrepancies were identified in the intermediate-confidence bins.

As shown in Table 3, the calibration analysis by confidence bins revealed heterogeneous local calibration patterns across probability intervals. Specifically, the 0.6–0.7 confidence interval showed an observed accuracy higher than the mean confidence, indicating a local tendency to underestimate predictive certainty. A similar pattern was observed in the 0.5–0.6 interval, where empirical accuracy exceeded the average predicted probability. In contrast, the 0.4–0.5 interval exhibited an observed accuracy substantially lower than the mean confidence, suggesting overconfidence in lower-certainty predictions. The smallest absolute gaps were observed in the 0.7–0.8 and 0.8–0.9 intervals, indicating comparatively better local calibration in higher-confidence predictions. Overall, these findings indicate that the model remains suitable for classification purposes; however, its predicted probabilities should be interpreted with caution when used to support threshold-based decision-making.

Table 3. Calibration by confidence bins.

Confidence Bin	<i>n</i>	Mean Confidence	Observed Accuracy	Absolute Gap
0.4–0.5	12	0.471	0.250	0.221
0.5–0.6	45	0.567	0.756	0.189
0.6–0.7	37	0.641	0.892	0.251
0.7–0.8	23	0.763	0.696	0.068
0.8–0.9	205	0.851	0.912	0.061

Note. The table summarizes calibration across confidence intervals. Mean confidence corresponds to the average predicted probability within each bin, observed accuracy represents the empirical proportion of correct predictions in that bin, and the absolute gap is the absolute difference between mean confidence and observed accuracy. Smaller absolute gaps indicate better local calibration.

Integrating explainability techniques, such as SHAP and LIME, with calibration analyses improves the interpretability of predictive outputs and supports their assessment in applied contexts. As shown in Figure 1, the SHAP beeswarm plot provides a global interpretation of the Class 4 one-vs-rest LightGBM model by ranking predictors according to their mean absolute contribution to the model output. These global feature importance patterns, together with the local explanations provided by LIME, allow model predictions to be decomposed into variable-level contributions. This facilitates model traceability and provides insight into how the trained classifier uses the available predictors to generate both individual and aggregate predictions. This approach enhances the transparency of the predictive system and supports the evaluation of the stability, consistency, and clinical plausibility of the model outputs. Importantly, these feature contributions should be interpreted as model-based predictive associations rather than as evidence of causal relationships.



Figure 1. SHAP beeswarm plot for Class 4 in the one-vs-rest multiclass LightGBM model. Note. Each point represents a test-set observation, and the horizontal axis shows the SHAP value on the log-odds scale for the Class 4 one-vs-rest classifier. Positive values increase the model output toward Class 4, whereas negative values reduce it. Feature colors indicate low-to-high feature values, and variables are ranked by mean absolute SHAP value, reflecting their relative contribution to the prediction. Clustered patterns may result from preprocessing procedures such as one-hot encoding or standardization and should not be interpreted as clinical thresholds.

The SHAP analysis showed that the Class 4 prediction was mainly driven by a small subset of variables, with frequency_of_consuming_vegetables, number_of_main_meals_daily, and age contributing the greatest overall influence. Lower vegetable consumption was consistently associated with positive SHAP values, indicating an increased model-estimated

likelihood of Class 4, whereas higher vegetable consumption contributed negatively. In contrast, higher values of number_of_main_meals_daily and age tended to shift the model output toward Class 4. Additional predictors, including smoking_no, liquid_intake_daily, overweight_obese_family_no, height, physical_exercise, consumption_of_fast_food_no, and schedule_dedicated_to_technology, showed smaller but non-negligible contributions, with effects varying across individual observations.

These findings indicate that the LightGBM model relied primarily on dietary behavior, meal frequency, age, smoking status, familial overweight/obesity background, and lifestyle-related variables to discriminate Class 4 from the other classes. Because SHAP values quantify feature contributions to the model output rather than causal effects, the observed patterns should be interpreted strictly as model-dependent predictive associations. Moreover, clustered or discontinuous SHAP distributions may reflect preprocessing procedures, such as one-hot encoding and standardization, and should not be interpreted as clinical thresholds. External validation is required before translating these attribution patterns into clinical or public health conclusions.

As shown in Figure 2, the SHAP waterfall plot provides a local explanation for an individual observation classified as Class 4 by the one-vs-rest multiclass LightGBM model. The model output is on the log-odds scale. The explanation begins at the baseline value, $E[f(x)] = -6.83$, representing the expected model output for Class 4 before accounting for the specific transformed feature values of the observation. The SHAP contributions are then added sequentially, showing how each predictor shifts the model output from the baseline toward the final prediction. In this case, the cumulative effect of the feature contributions yields a final model output of approximately $f(x) = 15.4$, corresponding to a predicted probability close to 1.000 for Class 4.

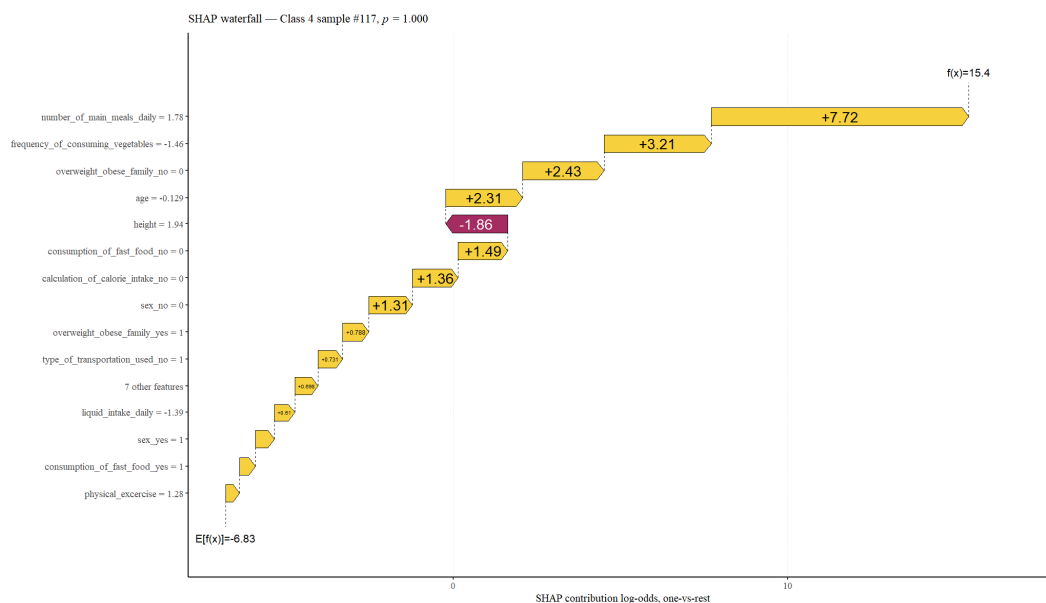


Figure 2. Local SHAP waterfall plot for the predicted class in the one-vs-rest multiclass LightGBM model. Note. The plot shows the local contribution of each feature to the prediction of a single observation. The explanation starts from the expected model output, $E[f(x)]$, which represents the baseline log-odds for the corresponding class. Each bar indicates the SHAP contribution of one feature: positive values increase the model output toward the predicted class, whereas negative values reduce it. The cumulative sum of the baseline value and feature contributions produces the final model output, $f(x)$, which can be converted into a predicted probability through the logistic function. Because categorical variables were one-hot encoded, apparent category indicators or cutoff-like values should be interpreted as model-processing artifacts rather than clinical thresholds.

For this observation, most local feature contributions were positive, indicating that the specific combination of transformed predictor values strongly supported assigning the observation to Class 4. The largest positive contribution came from `number_of_main_meals_daily`, followed by `frequency_of_consuming_vegetables`, `overweight_obese_family_no`, `age`, `consumption_of_fast_food_no`, and `calculation_of_calorie_intake_no`. These features increased the model-estimated log-odds for Class 4, shifting the prediction away from the baseline toward the obesity class. In contrast, height produced a negative SHAP contribution, reducing the estimated log-odds for Class 4 in this case. The remaining features had comparatively smaller local effects on the prediction.

This interpretation should be understood as local, predictive, and model-dependent. The SHAP values describe how the trained LightGBM classifier used the transformed representation of this specific observation to generate its prediction; they do not establish causal effects, biological mechanisms, or generalizable clinical rules. Likewise, variables derived from one-hot encoding, such as category-specific indicators, and standardized numerical values should not be interpreted as clinically meaningful thresholds. Broader conclusions about predictor relevance should be based on complementary global SHAP analyses, stability checks via repeated model refitting or bootstrapping, and external validation in independent populations.

As shown in Figure 3, LIME was used to provide a local approximation of the multiclass LightGBM prediction for an individual observation selected from the upper end of the holdout distribution's confidence range. Although the instance corresponded to the 85th percentile of the model-confidence distribution within the evaluated sample, the maximum calibrated probability remained moderate. In Figure 3a, the calibrated classifier assigned the highest probability to Class 2 ($p_{cal} = 48.4\%$), with non-negligible probabilities for the remaining classes. Therefore, this example should not be interpreted as a high-confidence prediction but rather as a case of local model traceability under classification uncertainty.

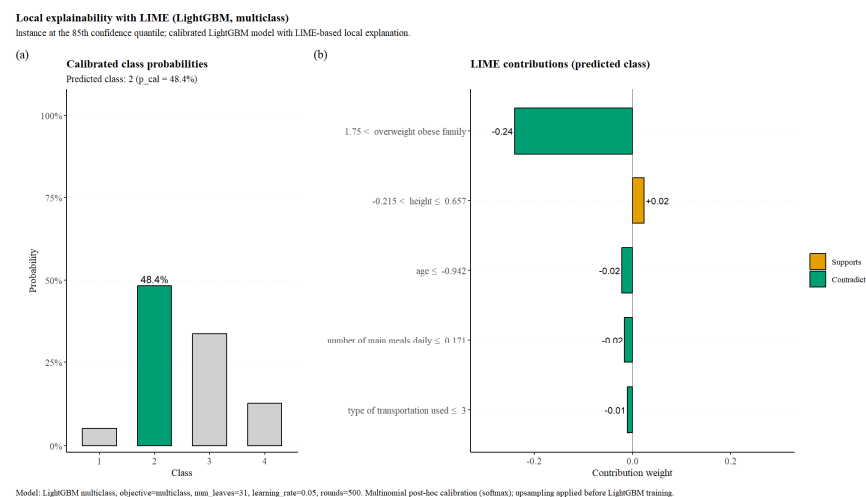


Figure 3. Local explainability with LIME for the multiclass LightGBM classifier. Note. Panel (a) shows the calibrated class probabilities for the analyzed observation, with the predicted class identified by the highest probability. Panel (b) presents the LIME-based local explanation for that prediction. Positive contribution weights support the predicted class, whereas negative weights oppose it. The displayed intervals and cutoff-like values result from LIME discretization and preprocessing of standardized predictors; therefore, they should be interpreted as model-explanation artifacts rather than clinical thresholds.

Figure 3b presents the LIME explanation for the predicted class. The largest local contribution was associated with family history of overweight/obesity, which had a negative weight and therefore contradicted the assignment to Class 2 for this instance. Smaller

positive or negative contributions were observed for height, age, number of main meals per day, and type of transportation used. These contribution weights describe how the trained LightGBM classifier used the feature values of this particular observation within the local neighborhood generated by LIME. Accordingly, they should be interpreted as local, model-dependent predictive contributions rather than as generalizable effects.

The inequalities and cutoff-like values shown in the LIME plot reflect the technical discretization of standardized predictors and do not correspond to clinically validated thresholds. Similarly, the sign and magnitude of the LIME weights indicate whether each feature locally supported or contradicted the predicted class, but they do not imply causal effects, clinical mechanisms, or population-level associations. This clarification is particularly important because the predicted class probability was moderate, indicating that the local explanation describes the model's decision process in an uncertain classification scenario.

To minimize interpretability bias from synthetic data, oversampling was applied only during model training, not during the generation of local explanations. SHAP and LIME explanations were computed using real observations from the test or holdout set, thereby preserving the original data distribution at the explanation stage. Nevertheless, because the classifier was trained on a rebalanced distribution, the learned decision boundaries—and consequently the resulting local attributions—may still have been indirectly influenced by the oversampling procedure.

Overall, the LIME results should be interpreted as exploratory, predictive, and model-dependent. Their purpose is to assess local traceability and the plausibility of the classifier's decision for an individual observation, rather than to establish causal or clinical mechanisms. In the analyzed instance, the Class 2 prediction resulted from the combined influence of several predictors, with one dominant contradictory contribution and several smaller supporting or opposing effects. Because the winning-class probability was only moderate, this case is best understood as an illustration of how LIME can help audit model behavior when class membership is uncertain. Further validation in independent datasets, together with study designs specifically intended for causal inference, would be required before drawing mechanistic or clinical conclusions.

5. Discussion

The present findings do not support the existence of a single classifier with statistically demonstrable superiority over the remaining models. Although the Friedman test identified significant global differences in weighted F1-score among the evaluated algorithms, the Nemenyi post hoc analysis showed that the best-performing methods formed a statistically comparable group. Specifically, Stacking, Random Forest, LightGBM, Voting, Gradient Boosting, and Extra Trees were assigned to Group A, indicating that their differences were not statistically significant after adjustment for multiple comparisons. Therefore, the observed ranking should be interpreted as evidence of relative performance trends rather than definitive statistical dominance.

This distinction is methodologically relevant. In nested cross-validation, model performance may vary across folds due to data heterogeneity, class imbalance, and model-specific sensitivity to training partitions. Consequently, average ranks provide a robust summary of relative behavior across resampling iterations, whereas Nemenyi groupings prevent the overinterpretation of small numerical differences. Under this inferential framework, Stacking showed the most favorable average rank, while LightGBM achieved the highest mean weighted F1-score. However, neither model demonstrated statistically significant superiority over the other top-tier classifiers. Thus, the empirical evidence supports the existence of a compact cluster of high-performing and statistically indistinguishable models rather than a single optimal classifier.

From an applied perspective, the selection of LightGBM as the final model can be justified on practical grounds. LightGBM combined the highest mean weighted F1-score with competitive ranking stability and membership in the best statistical group. This is consistent with previous evidence showing the effectiveness of gradient-boosted tree models in structured biomedical and obesity-related prediction tasks, including applications based on electronic health records and clinical cohorts [39,42]. Nevertheless, its selection should not be interpreted as proof of inferential superiority over Stacking, Random Forest, Voting, Gradient Boosting, or Extra Trees, but rather as a pragmatic decision supported by its balance between predictive performance, robustness, and implementation feasibility.

The strong performance of Stacking is also noteworthy. As a meta-ensemble, Stacking can exploit complementary predictive patterns among heterogeneous base learners, which may explain its favorable average rank. This finding is consistent with prior studies suggesting that hybrid or ensemble combination strategies can improve predictive stability by integrating models with diverse error profiles [43]. However, the absence of statistically significant differences within Group A indicates that the advantage of Stacking is relative rather than conclusive. In addition, its higher computational complexity and potentially weaker calibration profile should be considered before recommending it as the preferred option in operational settings.

The intermediate position of Bagging and XGBoost, both assigned to the overlapping A–B group, further reinforces the importance of cautious interpretation. Their performance was not clearly separated from either the highest-performing group or the lower-performing group, suggesting moderate but less stable competitiveness. In contrast, AdaBoost was the only classifier assigned exclusively to Group B, confirming its inferior relative performance for the multiclass obesity classification task. This result is consistent with the broader observation that more robust ensemble and boosted-tree approaches generally outperform less adaptive classical ensemble configurations in complex clinical prediction problems.

Beyond discrimination, clinical and public-health applications require models that are interpretable and probabilistically reliable. The use of explainability methods such as SHAP, TreeSHAP, and LIME is therefore essential to assess whether model predictions are traceable, clinically plausible, and consistent with known obesity-related predictors. Previous studies have emphasized the value of SHAP-based interpretation in LightGBM and other machine-learning models for obesity-status classification, particularly when translating model outputs into understandable risk profiles [39,40,46,47]. However, these explanations should be interpreted as model-dependent predictive associations and not as evidence of causal effects.

Calibration is another critical consideration. In clinical decision-support contexts, reliable probabilities are often more useful than class labels alone, particularly when predictions inform screening, prioritization, or risk stratification. Therefore, discrimination metrics such as weighted F1-score, accuracy, and ROC-AUC should be complemented with calibration measures, including the Brier score, expected calibration error, and reliability plots. This recommendation is aligned with prior methodological guidance emphasizing the joint reporting of discrimination and calibration in obesity and overweight prediction models [55]. Accordingly, even when a model shows strong classification performance, its practical utility depends on whether its predicted probabilities are sufficiently reliable for decision-making.

Overall, the results indicate that model selection should not be based solely on the highest rank or marginal differences in predictive metrics. Instead, when several classifiers are statistically comparable, the final choice should incorporate additional criteria such as calibration quality, interpretability, computational cost, robustness, and ease of deployment. In this study, LightGBM represents a well-supported practical choice, while Stacking

remains a highly competitive meta-ensemble alternative. Random Forest, Voting, Gradient Boosting, and Extra Trees also constitute viable options when implementation simplicity, stability, or interpretability are prioritized. Thus, the main contribution of the comparative analysis is not the identification of a universally superior model, but the demonstration of a statistically supported set of robust alternatives for multiclass obesity-status classification.

6. Conclusions

This study provides a comparative evaluation of ensemble and meta-ensemble learning methods for multiclass obesity-status classification using a leakage-aware validation framework. The statistical analysis showed significant global differences among classifiers; however, the Nemenyi post hoc test did not identify a single model with clear statistical superiority. Instead, the best-performing algorithms formed a statistically comparable group, indicating that model selection should be interpreted in terms of inferential equivalence rather than absolute dominance.

Within this statistically comparable group, LightGBM was selected as the final model because it achieved the highest mean weighted F1-score and showed a favorable overall profile across discrimination, agreement, and calibration-related measures. Nevertheless, this selection should be understood as a practical and methodological decision, rather than as evidence of statistically significant superiority over Stacking, Random Forest, Voting, Gradient Boosting, or Extra Trees. Stacking also showed competitive predictive performance, supporting the usefulness of meta-ensemble strategies; however, its higher computational complexity and weaker calibration profile should be considered before practical implementation.

The main limitations of this study include the use of a single publicly available dataset, the absence of independent external validation, the potential influence of oversampling on learned decision boundaries, and the fact that SHAP- and LIME-based explanations reflect model-dependent predictive associations rather than causal mechanisms. Moreover, because the dataset was derived from a specific population and data-collection context, the findings should not be directly generalized to other clinical or public-health settings without further validation.

Future research should evaluate these models in independent, multicenter, and demographically diverse cohorts; examine subgroup-level performance and fairness; assess calibration across different populations; analyze the stability of SHAP- and LIME-based explanations; and monitor model behavior under real-world deployment conditions. These additional analyses are necessary to determine whether the observed performance patterns remain consistent beyond the dataset analyzed in this study.

Author Contributions: Conceptualization, D.A.-G. and W.M.-R.; Methodology, D.A.-G. and W.M.-R.; Software, D.A.-G.; Validation, D.A.-G., A.P., E.O.-T. and F.B.-S.; Formal analysis, D.A.-G., W.M.-R. and A.P.; Investigation, D.A.-G., W.M.-R. and A.P.; Resources, A.P., E.O.-T. and F.B.-S.; Writing—original draft, D.A.-G. and W.M.-R.; Writing—review & editing, D.A.-G. and W.M.-R.; Visualization, D.A.-G., E.O.-T. and F.B.-S.; Supervision, W.M.-R.; Project administration, A.P., E.O.-T. and F.B.-S.; Funding acquisition, A.P., E.O.-T. and F.B.-S. 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. The dataset analyzed in this study was obtained from a publicly available Kaggle resource on obesity-status classification. Detailed information could be found in Section 3.

Informed Consent Statement: Patient consent was waived because this research was a retrospective study.

Data Availability Statement: The data utilized in this study were obtained from a publicly available dataset hosted on Kaggle (<https://www.kaggle.com/code/zachorwin/data-science-for-change#Introduction-and-Problem-Statement>) (accessed on 27 June 2025). No new data were created. The dataset is accessible to the research community through the Kaggle platform.

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

References

1. Mehmood, D.A. Epidemiological Trends and Risk Factors of Non-Communicable Diseases: A Global Perspective. *Non-Commun. Dis. Care* **2025**, *2*, 12–22.
2. Colmenarejo, G. Machine Learning Models to Predict Childhood and Adolescent Obesity: A Review. *Nutrients* **2020**, *12*, 2466. [CrossRef]
3. Ferdowsy, F.; Rahi, K.S.A.; Jabiullah, M.I.; Habib, M.T. A Machine Learning Approach for Obesity Risk Prediction. *Curr. Res. Behav. Sci.* **2021**, *2*, 100053. [CrossRef]
4. Vekic, J.; Stefanovic, A.; Zeljkovic, A. Obesity and Dyslipidemia: A Review of Current Evidence. *Curr. Obes. Rep.* **2023**, *12*, 207–222. [CrossRef]
5. Verma, M.; Kapoor, N.; Senapati, S.; Singh, O.; Bhadoria, A.S.; Khetarpal, P.; Kumar, S.; Bansal, K.; Ranjan, R.; Kakkar, R.; et al. Comprehending the Epidemiology and Aetiology of Childhood Obesity: Integrating Life Course Approaches for Prevention and Intervention. *Diabetes Ther.* **2025**, *16*, 1177–1206. [CrossRef]
6. Valenzuela, P.L.; Carrera-Bastos, P.; Castillo-García, A.; Lieberman, D.E.; Santos-Lozano, A.; Lucia, A. Obesity and the Risk of Cardiometabolic Diseases. *Nat. Rev. Cardiol.* **2023**, *20*, 475–494. [CrossRef]
7. Safaei, M.; Sundararajan, E.A.; Driss, M.; Boulila, W.; Shapi'i, A. A Systematic Literature Review on Obesity: Understanding the Causes & Consequences of Obesity and Reviewing Various Machine Learning Approaches Used to Predict Obesity. *Comput. Biol. Med.* **2021**, *136*, 104754. [CrossRef]
8. Wolfenden, L.; Ezzati, M.; Larijani, B.; Dietz, W. The Challenge for Global Health Systems in Preventing and Managing Obesity. *Obes. Rev.* **2019**, *20*, 185–193. [CrossRef]
9. Sørensen, T.I.A. Forecasting the Global Obesity Epidemic through 2050. *Lancet* **2025**, *405*, 756–757. [CrossRef]
10. Aremu, S.O.; Akute, B.; Aremu, D.O.; Zando, C.; Aremu, E.D.; Nwachukwu, O.J.; Omosibi, M.O.; Akute, V.O.; Oluwole, S.T.; Barkhadle, A.A.; et al. Dietary Strategies for Preventing and Managing Obesity through Evidence-Based Nutritional Interventions. *Discov. Public Health* **2025**, *22*, 424. [CrossRef]
11. Couto, F.D.F.S.; de Almeida, C.P.B. Mobile and Web Apps for Weight Management in Overweight and Obese Adults: An Updated Umbrella Review and Meta-Analysis. *Int. J. Environ. Res. Public Health* **2025**, *22*, 1152. [CrossRef]
12. Grossman, E. Obesity and Cardiovascular Disease: Mechanistic Insights and Nutritional Management Strategies. *Nutrients* **2025**, *17*, 2281. [CrossRef]
13. Lyn, R.; Heath, E.; Dubhashi, J. Global Implementation of Obesity Prevention Policies: A Review of Progress, Politics, and the Path Forward. *Curr. Obes. Rep.* **2019**, *8*, 504–516. [CrossRef]
14. Qian, F.; Wang, L. The Future of Obesity Management: Bridging Pharmacologic Innovation and Public Health. *Diabetes Metab. Syndr. Obes. Targets Ther.* **2025**, *18*, 4667–4670. [CrossRef] [PubMed]
15. Ehlers, L.H.; Reinstrup, N.W.; Olesen, R.H.; Holm, J.-C.; McEwan, P.; Le Roux, C.W. Global Barriers to Decision Makers for Prioritizing Interventions for Obesity. *Int. J. Obes.* **2025**, *49*, 246–253. [CrossRef] [PubMed]
16. Ganie, S.M.; Pramanik, P.K.D.; Zhao, Z. Ensemble Learning with Explainable AI for Improved Heart Disease Prediction Based on Multiple Datasets. *Sci. Rep.* **2025**, *15*, 13912. [CrossRef] [PubMed]
17. Airlangga, G. Machine Learning-Based Obesity Classification: A Comparative Study Using Self-Reported Survey Data and Ensemble Learning Models. *J. Teknol. Inform. Dan Komput.* **2025**, *11*, 347–361. [CrossRef]
18. Çizmecı, İ.H.; İncekara, H. Stacking Ensemble Based Hybrid Machine Learning Approach for Predicting Obesity Levels. In Proceedings of the 2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA), Ankara, Türkiye, 23–24 May 2025; pp. 1–8.
19. Jeong, J.-H.; Lee, I.-G.; Kim, S.-K.; Kam, T.-E.; Lee, S.-W.; Lee, E. DeepHealthNet: Adolescent Obesity Prediction System Based on a Deep Learning Framework. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 2282–2293. [CrossRef]
20. Barus, O.P.; Billie, F.; Pangaribuan, J.J.; Maulana, A. Obesity Prediction: K-Nearest Neighbor vs. Support Vector Machine. In Proceedings of the 2024 2nd International Conference on Technology Innovation and Its Applications (ICTIIA), Baku, Azerbaijan, 12–13 September 2024; pp. 1–5.
21. Sewpaul, R.; Awe, O.O.; Dogbey, D.M.; Sekgala, M.D.; Dukhi, N. Classification of Obesity among South African Female Adolescents: Comparative Analysis of Logistic Regression and Random Forest Algorithms. *Int. J. Environ. Res. Public Health* **2023**, *21*, 2. [CrossRef]

22. Ghosh, K.; Bellinger, C.; Corizzo, R.; Branco, P.; Krawczyk, B.; Japkowicz, N. The Class Imbalance Problem in Deep Learning. *Mach. Learn.* **2024**, *113*, 4845–4901. [\[CrossRef\]](#)
23. Barbierato, E.; Gatti, A. The Challenges of Machine Learning: A Critical Review. *Electronics* **2024**, *13*, 416. [\[CrossRef\]](#)
24. Präger, M.; Kurz, C.; Holle, R.; Maier, W.; Laxy, M. A Spatial Obesity Risk Score for Describing the Obesogenic Environment Using Kernel Density Estimation: Development and Parameter Variation. *BMC Med. Res. Methodol.* **2023**, *23*, 65. [\[CrossRef\]](#)
25. Danilova, M.; Dvurechensky, P.; Gasnikov, A.; Gorbunov, E.; Guminov, S.; Kamzolov, D.; Shibaev, I. Recent Theoretical Advances in Non-Convex Optimization. In *High-Dimensional Optimization and Probability: With a View Towards Data Science*; Nikeghbali, A., Pardalos, P.M., Raigorodskii, A.M., Rassias, M.T.H., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 79–163.
26. Keshavamurthy, D.; Kumar, M.; Tsaramiris, G.; Oroumchian, F. An AI-Based Framework for Secure and Transparent Banking: Integrating Adversarial Robustness, Interpretability, and Organizational Modeling. *Secur. Priv.* **2026**, *9*, e70153. [\[CrossRef\]](#)
27. Sharief, F.; Ijaz, H.; Shojafar, M.; Naeem, M.A. Multi-Class Imbalanced Data Handling with Concept Drift in Fog Computing: A Taxonomy, Review, and Future Directions. *ACM Comput. Surv.* **2024**, *57*, 1–48. [\[CrossRef\]](#)
28. Oeding, J.F.; Krych, A.J.; Pearle, A.D.; Kelly, B.T.; Kunze, K.N. Medical Imaging Applications Developed Using Artificial Intelligence Demonstrate High Internal Validity Yet Are Limited in Scope and Lack External Validation. *Arthrosc. J. Arthrosc. Relat. Surg.* **2025**, *41*, 455–472. [\[CrossRef\]](#)
29. Jin, D.; Sergeeva, E.; Weng, W.-H.; Chauhan, G.; Szolovits, P. Explainable Deep Learning in Healthcare: A Methodological Survey from an Attribution View. *WIREs Mech. Dis.* **2022**, *14*, e1548. [\[CrossRef\]](#)
30. Ayua, S.I. Random Forest Ensemble Machine Learning Model for Early Detection and Prediction of Weight Category. *J. Data Sci. Intell. Syst.* **2024**, *2*, 233–240. [\[CrossRef\]](#)
31. Azad, M.; Khan, M.F.K.; El-Ghany, S.A. XAI-Enhanced Machine Learning for Obesity Risk Classification: A Stacking Approach With LIME Explanations. *IEEE Access* **2025**, *13*, 13847–13865. [\[CrossRef\]](#)
32. Ramakrishna, M.T.; Venkatesan, V.K.; Izonin, I.; Havryliuk, M.; Bhat, C.R. Homogeneous Adaboost Ensemble Machine Learning Algorithms with Reduced Entropy on Balanced Data. *Entropy* **2023**, *25*, 245. [\[CrossRef\]](#)
33. Alqudah, A.M.; Moussavi, Z. Bridging Signal Intelligence and Clinical Insight: A Comprehensive Review of Feature Engineering, Model Interpretability, and Machine Learning in Biomedical Signal Analysis. *Appl. Sci.* **2025**, *15*, 12036. [\[CrossRef\]](#)
34. Alsulamy, S. Predicting Construction Delay Risks in Saudi Arabian Projects: A Comparative Analysis of CatBoost, XGBoost, and LGBM. *Expert Syst. Appl.* **2025**, *268*, 126268. [\[CrossRef\]](#)
35. Heddiam, S. Explainability of Machine Learning Using Shapley Additive exPlanations (SHAP): CatBoost, XGBoost and LightGBM for Total Dissolved Gas Prediction. In *Machine Learning and Granular Computing: A Synergistic Design Environment*; Pedrycz, W., Chen, S.-M., Eds.; Springer Nature: Cham, Switzerland, 2024; pp. 1–25.
36. Dugan, T.M.; Mukhopadhyay, S.; Carroll, A.; Downs, S. Machine Learning Techniques for Prediction of Early Childhood Obesity. *Appl. Clin. Inform.* **2015**, *6*, 506–520. [\[CrossRef\]](#)
37. Fu, L.; Wang, B.; Yuan, T.; Chen, X.; Ao, Y.; Fitzpatrick, T.; Li, P.; Zhou, Y.; Lin, Y.; Duan, Q.; et al. Clinical Characteristics of Coronavirus Disease 2019 (COVID-19) in China: A Systematic Review and Meta-Analysis. *J. Infect.* **2020**, *80*, 656–665. [\[CrossRef\]](#)
38. Shi, X.; Nikolic, G.; Epelde, G.; Arrúe, M.; Bidaurrezaga Van-Dierdonck, J.; Bilbao, R.; De Moor, B. An Ensemble-Based Feature Selection Framework to Select Risk Factors of Childhood Obesity for Policy Decision Making. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 222. [\[CrossRef\]](#)
39. Pang, G.; Shen, C.; Cao, L.; Hengel, A.V.D. Deep Learning for Anomaly Detection: A Review. *ACM Comput. Surv.* **2021**, *54*, 1–38. [\[CrossRef\]](#)
40. Cheng, E.R.; Steinhardt, R.; Ben Miled, Z. Predicting Childhood Obesity Using Machine Learning: Practical Considerations. *BioMedInformatics* **2022**, *2*, 184–203. [\[CrossRef\]](#)
41. Liu, B.; Zheng, D.; Zhou, S.; Chen, L.; Yang, J. VFDB 2022: A General Classification Scheme for Bacterial Virulence Factors. *Nucleic Acids Res.* **2022**, *50*, D912–D917. [\[CrossRef\]](#)
42. Solomon, D.D.; Khan, S.; Garg, S.; Gupta, G.; Almjally, A.; Alabdullah, B.I.; Alsagri, H.S.; Ibrahim, M.M.; Abdallah, A.M.A. Hybrid Majority Voting: Prediction and Classification Model for Obesity. *Diagnostics* **2023**, *13*, 2610. [\[CrossRef\]](#)
43. Abnoosian, K.; Farnoosh, R.; Behzadi, M.H. Prediction of Diabetes Disease Using an Ensemble of Machine Learning Multi-Classifiers Models. *BMC Bioinform.* **2023**, *24*, 337. [\[CrossRef\]](#)
44. Matt, J.E.; Rizzo, D.M.; Javed, A.; Eppstein, M.J.; Manukyan, V.; Gramling, C.; Dewoolkar, A.M.; Gramling, R. An Acoustical and Lexical Machine-Learning Pipeline to Identify Connectional Silences. *J. Palliat. Med.* **2023**, *26*, 1627–1633. [\[CrossRef\]](#)
45. Helforouh, Z.; Sayyad, H. Prediction and Classification of Obesity Risk Based on a Hybrid Metaheuristic Machine Learning Approach. *Front. Big Data* **2024**, *7*, 1469981. [\[CrossRef\]](#)
46. Sun, Z.; Yuan, Y.; Farrahi, V.; Herold, F.; Xia, Z.; Xiong, X.; Qiao, Z.; Shi, Y.; Yang, Y.; Qi, K.; et al. Using Interpretable Machine Learning Methods to Identify the Relative Importance of Lifestyle Factors for Overweight and Obesity in Adults: Pooled Evidence from CHNS and NHANES. *BMC Public Health* **2024**, *24*, 3034. [\[CrossRef\]](#)

47. Chatterjee, A.; Gerdes, M.W.; Martinez, S.G. Identification of Risk Factors Associated with Obesity and Overweight—A Machine Learning Overview. *Sensors* **2020**, *20*, 2734. [[CrossRef](#)]
48. Koklu, N.; Sulak, S.A. Using Artificial Intelligence Techniques for the Analysis of Obesity Status According to the Individuals' Social and Physical Activities. *Sinop. Univ. J. Nat. Sci.* **2024**, *9*, 217–239. [[CrossRef](#)]
49. Shahraki, A.; Abbasi, M.; Haugen, Ø. Boosting Algorithms for Network Intrusion Detection: A Comparative Evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. *Eng. Appl. Artif. Intell.* **2020**, *94*, 103770. [[CrossRef](#)]
50. Konstantinov, A.V.; Utkin, L.V. Interpretable Machine Learning with an Ensemble of Gradient Boosting Machines. *Knowl.-Based Syst.* **2021**, *222*, 106993. [[CrossRef](#)]
51. Jain, E.; Singh, A. Advanced Gradient Boosting Techniques for Predicting Obesity Risk: A Comprehensive Machine Learning Approach. In Proceedings of the 2024 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India, 11–13 December 2024; pp. 667–672.
52. Dagnew, G.; Shekar, B.H. Ensemble Learning-Based Classification of Microarray Cancer Data on Tree-Based Features. *Cogn. Comput. Syst.* **2021**, *3*, 48–60. [[CrossRef](#)]
53. Lian, W.; Nie, G.; Jia, B.; Shi, D.; Fan, Q.; Liang, Y. An Intrusion Detection Method Based on Decision Tree-Recursive Feature Elimination in Ensemble Learning. *Math. Probl. Eng.* **2020**, *2020*, 2835023. [[CrossRef](#)]
54. Trujillano, J.; Serviá, L.; Badia, M.; Serrano, J.C.E.; Bordejé-Laguna, M.L.; Lorencio, C.; Vaquerizo, C.; Flordelis-Lasierra, J.L.; de Lagrán, I.M.; Portugal-Rodríguez, E.; et al. Methodological Review of Classification Trees for Risk Stratification: An Application Example in the Obesity Paradox. *Nutrients* **2025**, *17*, 1903. [[CrossRef](#)]
55. Shaban, W.M.; El-Din Moustafa, H.; El-Seddek, M.M. Machine Learning Framework for Predicting Susceptibility to Obesity. *Sci. Rep.* **2025**, *15*, 35040. [[CrossRef](#)]

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.