Article

Not peer-reviewed version

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

Daniel Andrade-Girón , William Marin-Rodriguez * , Américo Peña , Elsa Oscuvilca-Tapia , Fredy Bermejo-Sanchez

Posted Date: 10 April 2026

doi: 10.20944/preprints202604.0697.v1

Keywords: obesity; risk stratification; machine learning; ensemble learning; stacking; random forest; gradient boosting; explainable AI

 doi:10.20944/preprints202604.0697.v1

Article

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

Daniel Andrade-Girón [1], William Marin-Rodriguez [2,*], Americo Peña [3], Elsa Oscuvilca-Tapia [4] and Fredy Bermejo-Sanchez [5]

[1] Department of Formal and Natural Sciences, Universidad Nacional José Faustino Sánchez Carrión, Huacho, Lima 15136, Peru
[2] Department of Engineering Systems, Computer and Electronics, Universidad Nacional José Faustino Sánchez Carrión, Huacho, Lima 15136, Perú
[3] Department of Medicine, Universidad Nacional José Faustino Sánchez Carrión, Huacho, Lima 15136, Peru
[4] Department of Nursing, Universidad Nacional José Faustino Sánchez Carrión, Huacho, Lima 15136, Peru
[5] Department of Medicine, Universidad Nacional José Faustino Sánchez Carrión, Huacho, Lima 15136, Peru
* Correspondence: wmarin@unjfsc.edu.pe; Tel.: +51990455214

Abstract

Obesity represents a significant public health concern, attributable to its high prevalence and its association with cardiometabolic comorbidities. This study compared a set of ensemble learning models—including canonical ensembles, meta-ensembles, and baselines for tabular data—in a multiclass obesity status prediction task using the "Obesity Dataset" (n = 1,610; 14 predictors; 4 classes). To ensure methodological rigor, a pipeline was implemented using ColumnTransformer, standardization, one-hot encoding, and rebalancing via SMOTENC applied exclusively to the training folds, thereby preventing data leakage. The performance of the system was evaluated using several evaluation metrics, including accuracy, F1-score, precision, recall, Cohen's kappa, and Matthews correlation coefficient. This evaluation was supplemented by a computational cost analysis. Inferential comparisons were executed using the Friedman test and the Nemenyi post-hoc test (α = 0.05). The findings indicated a high level of overall performance (≈89–90.5% precision), identifying a leading group of models that were statistically indistinguishable (Group A). This group included LightGBM (90.49% ± 1.38), Random Forest (90.16% ± 1.70), Stacking (90.21% ± 1.70), and Extra Trees (89.69% ± 1.55). It has been demonstrated that models such as XGBoost, Bagging, and CatBoost demonstrate competitive performance with partial statistical overlap. Conversely, Gradient Boosting and AdaBoost exhibited significantly lower performance. In summary, a single dominant model was not identified; rather, a set of equivalent solutions was identified. The selection of a model should be based on a balance between accuracy, computational cost, and interpretability. Random Forest and Extra Trees are efficient options, and Stacking is a valid alternative when maximizing predictive performance is prioritized.

Keywords: obesity; risk stratification; machine learning; ensemble learning; stacking; random forest; gradient boosting; explainable AI

1. Introduction

Obesity has emerged as a critical public health challenge on a global scale. This is due to three factors: its sustained growth, its complex etiology and pathogenesis, and its close association with chronic noncommunicable diseases [1]. These include type 2 diabetes mellitus [2], hypertension [3],

 doi:10.20944/preprints202604.0697.v1

dyslipidemia, and various types of cancer [4]. The impact of the disease is distributed across the entire life course, encompassing adults, adolescents, and children. Furthermore, the impact is heterogeneous across different countries and regions, reflecting variations in social, environmental, and behavioral determinants [5]. In this regard, obesity is not only a highly prevalent condition but also a central risk factor for the progression of cardiometabolic comorbidities and systemic complications [6].

Beyond the clinical harm it causes, obesity imposes a growing socioeconomic burden. Associated metabolic disorders, including insulin resistance [7], metabolic syndrome, low-grade chronic inflammation, and atherosclerotic cardiovascular disease [8], have been shown to increase the utilization of healthcare services, productivity losses, and healthcare expenditures, thereby compromising the sustainability of healthcare systems [9]. Projections indicate that prevalence will continue to rise over the coming decades [10]. According to recent estimates, more than 1 billion people are living with obesity, including approximately 650 million adults, 340 million adolescents, and 39 million children [11]. If current trends persist, the economic impact of overweight and obesity could reach USD 4.32 trillion annually by 2035 (≈3% of global GDP) [12], and by 2050, it is estimated that nearly 3.8 billion adults will be obese [13]. This scenario underscores the necessity for cost-effective approaches that prioritize prevention, early detection, and risk stratification.

In response to this crisis, numerous countries have implemented population-wide interventions, including taxes on sugary beverages, front-of-package labeling, regulations on food advertising, and community programs. These measures have yielded modest and often unsustainable benefits [14]. Concurrently, the most efficacious clinical strategies, including advanced pharmacotherapy and bariatric surgery, encounter substantial barriers related to cost, access, and scalability, particularly in low- and middle-income countries [15]. Indeed, an examination of the extant evidence suggests that the global response remains fragmented and insufficient, given the magnitude and speed of the problem [16], requiring a comprehensive, multisectoral, syndemic approach, complemented by analytical tools that enable the identification of at-risk subgroups and the prioritization of intervention actions with greater precision.

In this context, machine learning (ML) has emerged as a methodological approach with the potential to analyze clinical, biochemical, behavioral, and sociodemographic data, capturing nonlinear interactions and generating individualized predictions that support screening and prevention strategies [17]. Recent literature (2023–2025) demonstrates an augmentation in the utilization of predictive applications in the domain of cardiometabolic health, with a particular focus on models for tabular data and explainability frameworks designed to enhance clinical interpretability [18,19]. However, individual models—for example, artificial neural networks (ANNs) [20], support vector machines [21], and decision trees (DTs) [22]—have demonstrated significant limitations in real-world scenarios, characterized by heterogeneity, noise, uncertainty, and class imbalance [23]. The literature identifies several recurring problems, including overfitting [24], sensitivity to outliers [25], and optimization difficulties—including convergence to local optima in non-convex, high-dimensional spaces—which can result in unstable performance and reduced generalization ability [26].

The challenge, therefore, is not solely algorithmic; a methodological gap persists in the clinical application of ML to obesity. Specifically, there is an absence of standardized frameworks that comprehensively evaluate the robustness, traceability, and transparency of “black-box” models trained on heterogeneous data with class imbalance [27,28]. This phenomenon engenders bias when global metrics are accorded precedence over more informative indicators for multiclass and/or imbalanced classification, such as F1-score, area under the curve (AUC)-ROC, or per-class metrics [29]. This limitation is further compounded by the dearth of external validation [30], which curtails the transferability of these models to novel population contexts. Additionally, the incorporation of interpretability methods based on feature attribution—for example, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) is still in its early stages. This makes it challenging to systematically link the contribution of variables to clinical knowledge,

 doi:10.20944/preprints202604.0697.v1

detect inconsistencies, and audit potential biases. As a result, uncertainty remains regarding the actual utility of the models in public health settings [31].

In light of these limitations, ensemble learning approaches have emerged as effective alternatives for enhancing stability and predictive performance [32]. The integration of multiple classifiers, employing methodologies such as bagging, boosting, or stacking [33], enables ensemble models to regulate variance, mitigate bias, and enhance robustness in scenarios where data are incomplete, heterogeneous, or noisy [34]. Furthermore, several ensembles based on randomized trees offer advantages in terms of interpretability, a key aspect for their clinical acceptance and eventual implementation in public health programs [35]. In the context of contemporary tabular data, contemporary baselines such as XGBoost, LightGBM, and CatBoost are imperative competitive benchmarks due to their capacity to model nonlinearities and interactions, manage heterogeneity, and sustain high performance under class imbalance. Consequently, they must be incorporated into methodologically equitable comparisons [36,37].

Within this framework, the objective of this study is to develop and evaluate a set of canonical ensembles (e.g., Random Forest, Extra Trees, and AdaBoost/Gradient Boosting) and Stacking-based meta-ensembles for a multiclass classification scenario for obesity risk/status, comparing their performance using stratified cross-validation. The performance of these models is quantified using several metrics, including accuracy, precision, recall, Cohen's kappa (Cohen's κ), and Matthews correlation coefficient (MCC). All metrics are reported as the mean (%) ± standard deviation (SD) across the cross-validation folds, while computational time is expressed in seconds (s). This approach enables a comprehensive characterization of the discriminative power, performance stability, and degree of agreement of the evaluated models.

To support the inferential robustness of the benchmarking, a comparative framework will be used that includes nonparametric analysis (Friedman test) and Nemenyi post-hoc tests, enabling the identification of significant differences between models. Furthermore, agreement metrics (Cohen's κ and MCC) will be incorporated to quantify predictive consistency and alignment with the diagnostic reference. Additionally, global interpretability (SHAP) and local explainability (LIME) will be integrated to assess clinical coherence, identify determining variables, and explore potential bias patterns.

In accordance with the preceding framework, the study proposes a predictive, explainable, and interpretable architecture. This study offers three principal contributions, providing verifiable evidence for the aforementioned assertions: (i) it proposes a reproducible methodological protocol for the multiclass prediction of obesity risk/status that integrates preprocessing, handling of heterogeneity and data imbalance, hyperparameter optimization, and stratified cross-validation; (ii) it establishes a comparative benchmark against canonical ensembles and contemporary baselines based on Gradient Boosting (XGBoost, LightGBM, and CatBoost), supported by statistical inference using Friedman and Nemenyi post-hoc tests, and incorporating complementary metrics—MCC and Cohen's κ—to quantify stability, reliability, and agreement in performance; and (iii) it incorporates an integrated global–local interpretability framework—combining SHAP and LIME—to audit the model's clinical consistency, identify key predictors at the population level, and explain individual decisions, facilitating its transfer to applied scenarios for screening and clinical decision support.

2. Literature Review

Obesity represents a significant challenge to contemporary public health, owing to its multifactorial and multisystemic etiology. In this context, ML algorithms have demonstrated their efficacy as a robust methodological alternative for early risk prediction. These algorithms enable nonlinear modeling, capture high-order interactions, and integrate high-dimensional data with collinearity, noise, and class imbalance. By doing so, they overcome the limitations of traditional statistical approaches and facilitate clinically useful risk stratification.

One of these early studies is that of Dugan et al. [38], which presented an early prediction of childhood obesity before the age of two by analyzing clinical cohorts and classical algorithms (ID3,

 doi:10.20944/preprints202604.0697.v1

Random Forest, Naïve Bayes). First, although the study was innovative, it did not provide a comparative statistical analysis, and only conventional measures were evaluated.

A marked increase in methodological complexity has been observed since 2020. Fu et al. [39] incorporated interpretable modeling techniques (LightGBM and SHAP analysis) into a cohort of preterm infants while examining early body mass index (BMI) trajectories, along with maternal BMI, as important predictive factors. Concurrently, Colmenarejo [2] has voiced concerns regarding the application of ML and deep learning in the context of pediatric electronic health records (EHRs), concluding that these tools demonstrate superior performance in comparison to classical statistical regression methods. Furthermore, the author has identified deficiencies in the clinical validation process.

Concurrently, Ferdowsy et al. [3], reported that the Logistic Regression algorithm attained the highest level of accuracy (97.09%) in comparison to the other classifiers evaluated. Conversely, the Gradient Boosting algorithm exhibited the least optimal performance, with an accuracy of 64.08%, and also demonstrated the lowest values across the other metrics analyzed. The approaches underwent diversification. Shi et al. [40] applied the Bagging-based Feature Selection framework integrating MapReduce (BFSMR), an ensemble learning model combined with weighted feature voting. This approach was used to identify informative clinical factors among more than 426,000 patients. Pang et al. [41] applied XGBoost to a dataset comprising over 800,000 pediatric visits and achieved an AUC of 0.81 in their prediction of future obesity. In local areas, Ferdowsy et al. [3] provided complementary evidence, while Safaei et al. [7] conducted a systematic review of 93 studies. Notwithstanding their magnitude, these studies did not take into account adjusted p-values or multiple statistical comparisons.

In 2022, research concentrated on advanced architectures. Cheng et al. [8] employed long short-term memory (LSTM) networks to predict BMI in pediatric cohorts with missing values, and Liu et al. [42] utilized deep reinforcement learning to identify gut microbiome biomarkers for obesity.

Recent studies conducted in 2023 offer additional evidence that lends further credence to the efficacy of ensemble models. Solomon et al. [43] proposed a hybrid majority-voting model composed of Gradient Boosting and XGBoost phylogenetic tree algorithms, as well as multi-layer perceptron (MLP) ANNs, achieving 97.16% accuracy on intensive care unit (ICU) datasets. In a similar vein, Abnoosian et al. [44] employed a set of multiple classifiers for patients with diabetes, achieving an AUC close to 1.0. Additionally, Matt et al. [45] conducted a comparative analysis of 10 algorithms, determining that Random Forest and Logistic Regression emerged as the most resilient. However, the validity of these studies is constrained by the absence of multicenter external validation through rigorous statistical analysis to ascertain the statistical significance of the observed differences between algorithms.

Recent research has centered on the development of hybrid models and the integration of interpretability techniques. Helforoush and Sayyad [46] presented a novel approach that integrates an ANN-Please check whether the included term expansion is appropriate: particle swarm optimization (PSO**)** method, enhanced by the use of SHAP. This approach has been shown to yield significant advancements in comparison to conventional regression methods. In a similar vein, Sun et al. [47] employed Gradient Boosted Decision Tree (GBDT) on large international cohorts (China Health and Nutrition Survey [CHNS] and National Health and Nutrition Examination Survey [NHANES]) to identify physical inactivity and alcohol consumption as key determinants in predicting obesity. In a similar vein, Airlangga [18] investigated the potential of obesity classification through the utilization of self-reported data from Mexico, Peru, and Colombia. This study involved the evaluation of eight ML algorithms. The study demonstrated the effectiveness of ensemble techniques, particularly Gradient Boosting, with an accuracy of 96.49%, in capturing complex interactions among demographic, behavioral, and lifestyle variables. It also highlighted the limitations of simpler classifiers, such as Naïve Bayes and AdaBoost. Conversely, Çizmeci and İncekara [58] developed a hybrid model based on ensemble Stacking that achieved 98.58% accuracy, with consistent metrics in terms of F1-score, AUC, and a κ-value of 0.9834. This model demonstrated

 doi:10.20944/preprints202604.0697.v1

superior performance in comparison to individual classifiers and exhibited considerable promise for incorporation into public health decision-support systems. Consequently, it provided substantial evidence supporting the efficacy of hybrid and ensemble approaches in obesity risk classification.

In summary, the reviewed literature demonstrates a persistent transition from classical approaches toward more complex architectures, accompanied by documented enhancements in accuracy, robustness, and generalizability. However, this evolution is accompanied by a methodological gap that limits the comparability and inferential scope of the results. First, there is a largely unsystematic incorporation of adjusted agreement metrics—in particular, Cohen's κ and MCC—despite their utility for evaluating performance in multilabel contexts and potential imbalance. Indeed, these metrics are explicitly reported only in the study by Çizmeci and İncekara [19] Second, the application of nonparametric statistical procedures (e.g., Friedman with Nemenyi post-hoc comparisons) to rigorously establish whether the observed differences between algorithms—including contemporary Gradient Boosting-based baselines—are statistically significant under a comparable evaluation protocol is uncommon.

Moreover, there is a paucity of systematic analyses of interpretability and explainability employing model-agnostic approaches, such as SHAP for global interpretation and LIME for local explanation of predictions. The dual absence of sufficient inferential support and constrained explanatory auditing impedes the generalizability of findings, hinders the critical evaluation of clinical coherence, and undermines the formulation of standardized methodological recommendations. Consequently, the development of a benchmarking framework is imperative. This framework should not only facilitate comparisons between models and competitive benchmarks, such as XGBoost, LightGBM, and CatBoost, but also incorporate methods such as Friedman–Nemenyi, Cohen's κ/MCC, and a SHAP/LIME scheme. This integrated approach will empower researchers to draw comparative and clinically interpretable conclusions, thereby enhancing the understanding of the field.

3. Methodology

The development of the model was informed by the methodology employed in ML design, which is outlined below.

3.1. Data Source

A publicly available dataset on Kaggle was utilized from the "Obesity Dataset" repository (Sulak, 2024; accessed June 27, 2025; URL: https://www.kaggle.com/code/zachorwin/data-science-for-change#Introduction-and-Problem-Statement), whose source, collection protocol, and structure are documented in the source article by Köklü and Sulak published in the *Sinop University Journal of Natural Sciences* (2024). In the aforementioned study, data were collected via an online survey administered to 1,610 participants residing in Turkey, with institutional ethical approval (decision 2023/201, Necmettin Erbakan University). For this study, the version with 15 variables was utilized. The model comprises 14 predictors and 1 target variable, designated as "Class." The predictor variables in question were as follows: 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; while the target variable (Class) defines a multiclass outcome of body weight status into four categories coded as 1–4 (Underweight = 1, Normal = 2, Overweight = 3, Obesity = 4), with distributions of 73 (4.5%), 658 (40.9%), 592 (36.8%), and 287 (17.8%), respectively. In conclusion, given that the source article has been published under a Creative Commons Attribution–NonCommercial 4.0 (CC BY-NC 4.0) license, the dataset was reused in accordance with the terms of that license. This ensured proper attribution and restricted its use to noncommercial purposes. Consequently, access to and utilization of the data were executed in accordance with the stipulated

 doi:10.20944/preprints202604.0697.v1

terms of use of the repository on Kaggle, encompassing the prevailing conditions for downloading, employing, and redistributing the data.

3.2. Data Preprocessing

An initial exploratory analysis of the dataset was conducted using descriptive statistics to identify missing values, data type inconsistencies, and potential anomalies in the distribution of variables relevant to obesity prediction. This analysis verified the structural integrity of the database, revealing no missing values, which ensured the completeness of the dataset from the outset.

As no missing values were identified, the application of imputation techniques was rendered unnecessary. To address the presence of outliers, Z-score analysis was employed to retain only those records that were plausible within the context of the risk prediction model and/or obesity levels. Furthermore, the drop_duplicates() function was employed to eliminate redundant observations, thereby preventing potential biases in the training process and ensuring the accuracy of predictive performance estimates.

Categorical variables were transformed using one-hot encoding with the OneHotEncoder and pandas.get_dummies() functions, ensuring they were properly represented in ML algorithms. Furthermore, normalization and standardization techniques—specifically StandardScaler—were employed to standardize scales and enhance numerical stability during model optimization.

3.2.1. Partitioning Protocol, Hyperparameter Tuning, Preprocessing, and Handling of Imbalance

To obtain an unbiased estimate of out-of-sample performance and prevent information leakage, a two-stage protocol was adopted. Initially, the complete dataset was partitioned using stratified sampling, allocating 80% to the development set and 20% to the independent test set. A random seed (random_state = 42) was employed to ensure reproducibility. The test set remained fixed during the training, preprocessing, balancing, and hyperparameter selection processes, ensuring it did not influence any modeling decisions.

Second, hyperparameter tuning and model selection were performed exclusively on the development set using stratified cross-validation (k = 3) implemented within RandomizedSearchCV. The F1-score was used as the primary metric due to its suitability for multiclass classification with potential class imbalance. To guarantee the integrity of the data and forestall cross-validation contamination, a preprocessing pipeline was implemented. This pipeline entailed the integration of a ColumnTransformer (one-hot encoding for categorical variables and standardization for numerical variables) into an imblearn Pipeline. The imblearn Pipeline comprises the balancing method and the classifier. In this design, the parameters of one-hot encoding, scaling, and SMOTENC are optimized solely on the training data of each fold and subsequently applied to the validation fold without undergoing additional retraining. This approach guarantees that the validation (and testing) sets remain uncontaminated by information leakage.

3.3. Selection of a Learning Algorithm

The selection of ensemble learning models for obesity prediction is based on their ability to increase the system's accuracy and robustness by integrating multiple base classifiers within a unified framework. This approach is particularly relevant in clinical and population-based settings characterized by high heterogeneity and the presence of nonlinear and high-order relationships among demographic, anthropometric, family, and behavioral variables that influence body weight and the risk of obesity.

In this context, the selected algorithms are particularly useful for capturing synergies and complex interactions among predictors, even in the presence of noisy or partially collinear data. This facilitates the development of generalizable and clinically relevant models for potential integration into public health decision support systems aimed at detecting and stratifying obesity risk.

 doi:10.20944/preprints202604.0697.v1

3.3.1. AdaBoost Classifier

The Adaptive Boosting (AdaBoost) algorithm is a boosting-based ensemble learning technique that combines multiple weak classifiers to form a more robust and accurate predictive model for obesity classification tasks [48]. The operation of the system is centered on the iterative allocation of weights to observations, with the objective of augmenting the weight of misclassified instances in each cycle. This approach ensures that subsequent classifiers are oriented toward identifying and rectifying the errors made [49]. The contribution of each classifier to the final model is determined inversely proportional to its error rate, favoring the most accurate estimators in the joint decision [50].

3.3.2. Gradient Boosting Classifier

The Gradient Boosting Classifier is an ensemble learning algorithm that sequentially builds additive models, progressively minimizing a loss function using gradient descent in the functional space [51]. In the context of obesity prediction tasks, the incorporation of a new base classifier into the system entails the correction of residual errors from previous models. This iterative process serves to progressively enhance the system's predictive capacity. In contrast to AdaBoost, which adjusts the weights of the samples based on their error, Gradient Boosting constructs a sequential model by adjusting each new estimator to the residual gradients, offering flexibility through different loss functions for classification [52].

3.3.3. Random Forest Classifier

The Random Forest Classifier algorithm was implemented, an ensemble learning method based on DTs that combines multiple classifiers built on random subsets of data and features [53]. In the context of obesity prediction, this approach enhances accuracy, reduces variability, and mitigates the risk of overfitting. The implementation was executed using the sklearn.ensemble module, incorporating hyperparameter tuning via grid search and stratified cross-validation [54].

3.3.4. Extra Trees Classifier

The Extra Trees Classifier (Extremely Randomized Trees) is a variant of Random Forest that introduces a higher degree of randomization into the construction of trees [55]. In contrast to the Random Forest method, which determines optimal split thresholds, the Extra Trees method assigns these thresholds randomly within the subset of features at each node. In the context of obesity prediction, this strategy has been shown to enhance diversity among trees, reduce variance, and improve stability against noise and overfitting [56].

3.3.5. Bagging Classifier

The Bagging Classifier enhances stability and accuracy by reducing variance through training on multiple subsets generated by bootstrap sampling [57]. In the context of obesity classification and prediction, the ultimate prediction is derived through majority voting, a method that mitigates the risk of overfitting and enhances the generalizability of the model.

3.4. Performance Metrics

The model's performance is evaluated using metrics derived from the confusion matrix, which summarizes the agreement between true labels and predictions in obesity classification tasks. As errors may manifest as false positives or false negatives, the following terms are defined: true positive (TP), false negative (FN), false positive (FP), and true negative (TN). Key metrics are derived from these values.

 doi:10.20944/preprints202604.0697.v1

3.4.1. Accuracy

The proportion of accurate predictions derived from the total number of observations. In obesity prediction, accuracy may be insufficient if there is class imbalance or if minimizing FNs is prioritized.

3.4.2. Precision

Accuracy of positive predictions; controls for FPs (e.g., classifying someone as obese who is not).

3.4.3. Sensitivity (Recall)

In the context of obesity prediction, this model quantifies the ability to correctly identify individuals with obesity (the positive class). An FN diagnosis can result in the failure to identify a genuine case, consequently delaying the implementation of preventive or therapeutic measures.

3.4.4. Specificity

The capacity to precisely ascertain individuals who are not obese, thereby reducing overdiagnosis and unnecessary interventions, is paramount.

3.4.5. F1-Score

This metric has proven to be a valuable tool in the analysis of unbalanced scenarios frequently observed in obesity prediction models. It integrates precision and recall, though in clinical practice, recall is often prioritized to minimize FN results.

3.5. Statistical Tests

To ensure valid comparisons among multiple obesity prediction algorithms, it is essential not to interpret differences in performance solely based on averages. Therefore, nonparametric rank-based tests are used across multiple cross-validation partitions and/or multiple datasets.

3.5.1. Friedman Test

The nonparametric Friedman test will be used to compare the performance of three or more models evaluated under the same partitions (e.g., stratified k-fold) and/or across multiple datasets. This will allow for the determination of whether there are overall differences in the ranges of relevant metrics (e.g., AUC, F1-score, sensitivity, or κ).

3.5.2. Nemenyi Post-Hoc Test

In the event that the Friedman test yields a significant result, Nemenyi post-hoc analysis will be implemented to identify which pairs of models differ statistically significantly. This analysis will control for the error of multiple comparisons using the critical difference based on the average ranks.

3.6. Interpretability and Explainability

Probabilistic calibration and local explainability procedures are incorporated to improve the interpretability and reliability of the model's outputs. Class probabilities are calibrated using multinomial regression (softmax) applied to probabilities previously regularized via Jeffreys smoothing (α = 0.5), with the aim of correcting overconfidence biases and ensuring probabilistic consistency. This methodological approach has been demonstrated to yield more precise probability estimates, thereby enhancing their interpretability and facilitating their utilization in decision-making scenarios.

 doi:10.20944/preprints202604.0697.v1

4. Results

The application of ensemble models to a heterogeneous clinical-epidemiological dataset revealed varying performance in obesity risk classification. A systematic comparison of algorithms in the presence of complex patterns of multimorbidity and lifestyle behaviors demonstrated a consistent advantage of ensemble approaches over individual classifiers, particularly in terms of discrimination stability and the identification of TPs under class-imbalanced distributions.

To ensure data integrity and reproducibility, a standardized protocol was implemented for the following 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. During the quality control stage, the completeness of the dataset was verified, confirming the absence of missing values; consequently, imputation was unnecessary. Subsequently, records were deduplicated to mitigate overfitting associated with pattern replication. Categorical or ordinal variables were transformed using one-hot encoding (excluding the first category to avoid perfect collinearity), while numerical variables were standardized with StandardScaler to homogenize scales and promote the stability/convergence of models sensitive to the magnitude of the variables. Finally, class imbalance was addressed using SMOTENC, applied exclusively to the training set following the train/test split, with the aim of preventing information leakage and preserving a valid estimate of out-of-sample performance. This workflow enabled the management of prevalent sources of bias in predictive studies based on demographic, anthropometric, familial, and behavioral variables, thereby enhancing the comparability across algorithms.

Modeling and metrics. A total of five canonical ensembles were evaluated: AdaBoost, Gradient Boosting, Random Forest, Extra Trees, and Bagging. In addition, two meta-ensembles were assessed: Stacking and Voting. To contextualize performance relative to contemporary benchmarks on tabular data, the comparison also included gradient boosting-based baselines: XGBoost, LightGBM, and CatBoost. The performance of the model was evaluated using several quantitative metrics, including accuracy, F1-score, precision, recall, Cohen's κ, and MCC. This approach ensured that the overall accuracy of the model was supported by metrics that were robust to class imbalance and sensitive to agreement that was not attributable to chance.

Hyperparameter optimization was performed using RandomizedSearchCV with stratified cross-validation across three partitions. In cases where appropriate, up to 10 configurations per algorithm were evaluated. The combination that maximized the study's target metric was selected based on inter-partition stability criteria. In bagging-based methods, both Random Forest and Extra Trees converged to configurations with a high number of trees, a loose minimum split threshold, and controlled depth (n_estimators = 500, min_samples_split = 2, max_depth = 20). This was consistent with an effective reduction in variance through averaging without a disproportionate increase in bias. In the boosting approach, Gradient Boosting achieved its optimal performance with a conservative learning regime offset by a substantial number and moderate depth of trees (n_estimators = 500, max_depth = 10, learning_rate = 0.05), favoring the capture of nonlinear interactions while maintaining stability. Concurrently, AdaBoost was optimized with a more aggressive update scheme and a reduced need for extensive aggregation (n_estimators = 100, learning_rate = 1.0).

In second-level ensembles, Stacking selected a Logistic Regression with regularization as its base estimator (final_estimator__C = 1.0), while the Voting scheme performed better in hard mode (voting = "hard"), suggesting greater discrimination in contexts of disagreement among classifiers. In both cases, the search space was bounded, so the random search effectively evaluated all available combinations. This circumstance was taken into account when interpreting the stability of the optimum.

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

 doi:10.20944/preprints202604.0697.v1

generalizability. This analysis provides a solid foundation for determining optimal configurations in practical applications and future lines of research.

Table 1. Overall predictive performance (mean (%) ± SD) and computational time.

Model	Acc. (%)	F1 (%)	Prec. (%)	Rec. (%)	κ (Cohen)	MCC	Time (s)
AdaBoost	75.62 ± 3.42	75.43 ± 3.39	75.72 ± 3.32	75.63 ± 3.43	0.675 ± 0.046	0.676 ± 0.045	10.97
Gradient Boosting	90.16 ± 2.24	90.14 ± 2.26	90.25 ± 2.27	90.16 ± 2.24	0.869 ± 0.030	0.869 ± 0.030	371.77
Random Forest	90.87 ± 2.39	90.82 ± 2.44	90.87 ± 2.40	90.88 ± 2.38	0.878 ± 0.032	0.879 ± 0.032	16.73
Extra Trees	89.83 ± 1.90	89.80 ± 1.97	89.86 ± 1.96	89.84 ± 1.89	0.864 ± 0.025	0.865 ± 0.025	14.81
Bagging	89.26 ± 2.66	89.19 ± 2.70	89.27 ± 2.65	89.27 ± 2.65	0.857 ± 0.035	0.857 ± 0.035	3.11
Stacking	91.25 ± 1.73	91.24 ± 1.75	91.28 ± 1.74	91.26 ± 1.73	0.883 ± 0.023	0.884 ± 0.023	2153.11
Voting	90.87 ± 2.05	90.84 ± 2.07	90.91 ± 2.06	90.87 ± 2.05	0.878 ± 0.027	0.879 ± 0.027	410.56
XGBoost	89.45 ± 2.38	89.40 ± 2.42	89.49 ± 2.41	89.45 ± 2.38	0.859 ± 0.032	0.860 ± 0.032	13.94
LightGBM	90.78 ± 2.39	90.74 ± 2.43	90.87 ± 2.39	90.78 ± 2.39	0.877 ± 0.032	0.878 ± 0.032	177.77
CatBoost	89.16 ± 2.63	89.08 ± 2.67	89.13 ± 2.66	89.16 ± 2.63	0.855 ± 0.035	0.856 ± 0.035	48.42

Note: Acc.=accuracy; Prec.=precision; Rec.=recall; κ=Cohen's kappa; MCC=Matthews correlation coefficient. All values are reported as mean (%) ± SD across cross-validation folds. Computational time is expressed in seconds (s).

A comparative evaluation of ensemble learning algorithms for obesity prediction revealed high and consistent overall performance across discrimination and agreement metrics. Ensemble-based models demonstrated consistent performance, with accuracy, F1-score, precision, and recall values ranging from 89 to 91%, indicating adequate generalization ability for the multiclass classification problem.

A comparison of the extensively utilized boosting models (XGBoost, LightGBM, and CatBoost) with the best meta-ensemble reveals that they attain competitive performance, though they do not surpass the latter. Specifically, LightGBM achieved an accuracy of 90.78% ± 2.39 (F1 = 90.74% ± 2.43; κ = 0.877 ± 0.032; MCC = 0.878 ± 0.032), followed by XGBoost (accuracy = 89.45% ± 2.38; κ = 0.859 ± 0.032) and CatBoost (accuracy = 89.16% ± 2.63; κ = 0.855 ± 0.035). These results indicate that, while boosting methods are considered robust baselines, their performance exhibits slight inferiority to that of more complex ensemble strategies.

In this particular context, the Stacking model was identified as the most optimal method, demonstrating the highest levels of performance across all key metrics. These metrics include accuracy of 91.25% ± 1.73, F1-score of 91.24% ± 1.75, precision of 91.28% ± 1.74, and recall of 91.26% ± 1.73. The study demonstrated the highest levels of agreement (κ = 0.883 ± 0.023; MCC = 0.884 ± 0.023), indicating substantial to high levels of agreement between predictions and actual values, with robust

 doi:10.20944/preprints202604.0697.v1

resistance to randomness. This outcome aligns with the extant literature, wherein meta-ensembles are known to frequently capture nonlinear relationships and complementarities among base models.

Random Forest (accuracy = 90.87% ± 2.39; κ = 0.878 ± 0.032), Voting (accuracy = 90.87% ± 2.05; κ = 0.878 ± 0.027), and Extra Trees (accuracy = 89.83% ± 1.90; κ = 0.864 ± 0.025), with accuracy differences of less than 1 percentage point (pp) compared to the best model. Gradient Boosting also showed solid performance (accuracy = 90.16% ± 2.24; κ = 0.869 ± 0.030), establishing itself as a competitive alternative. For its part, Bagging achieved slightly lower results (accuracy = 89.26% ± 2.66; κ = 0.857 ± 0.035), although still within an acceptable range of performance.

Conversely, AdaBoost exhibited the least optimal performance (accuracy = 75.62% ± 3.42; κ = 0.675 ± 0.046; MCC = 0.676 ± 0.045), indicating its inadequacy in discerning the underlying patterns in this problem, a deficiency that may be attributable to its susceptibility to noise and intricate class distributions.

The computational cost analysis yielded a discernible discrepancy in efficiency. Bagging was the most efficient method (3.11 s), followed by AdaBoost (10.97 s), XGBoost (13.94 s), Extra Trees (14.81 s), and Random Forest (16.73 s), all of which combine short processing times with high predictive performance. CatBoost (48.42 s) and LightGBM (177.77 s) fell into an intermediate category, while Gradient Boosting (371.77 s) and Voting (410.56 s) substantially increased computational cost. It is noteworthy that the Stacking model exhibited the longest execution time (2153.11 s), which significantly exceeded the execution times of the other models.

When considered as a whole, these results allow for the identification of a clear trade-off between performance and cost. While the Stacking method enhances the predictive power and agreement, it does so at the expense of a high computational cost, which limits its applicability in real-time scenarios. Conversely, models such as Random Forest and Extra Trees exhibit a notably advantageous performance–efficiency ratio, exhibiting only marginal losses in accuracy compared to the optimal model, while concurrently demonstrating substantial reductions in computation time.

From a clinical-operational perspective, the high degree of correlation between κ and MCC in the highest-performing models (≈0.86–0.88) suggests stability in classification even in the presence of potential class imbalances, which is critical in screening and risk stratification applications. In this regard, Random Forest and Extra Trees emerge as preferred candidates for implementation in environments with computational constraints, while Stacking stands out as an optimal solution when the goal is to maximize predictive performance in offline or batch processing contexts.

In summary, meta-ensembles are establishing themselves as the gold standard for obesity prediction. Nevertheless, the selection of a model must be predicated on an explicit balance between performance and computational cost. To that end, Stacking should be regarded as the option offering maximum accuracy, while Random Forest or Extra Trees should be considered as robust, efficient, and more viable alternatives for practical implementation. These findings are consistently reflected in the results summarized in Table 2, which details the statistical comparison of the evaluated *ensemble learning* models.

Table 2. Ranking of ensemble models for predicting obesity. The average accuracy is reported as the average ranks are derived from the Friedman test, and the statistical groups correspond to the Nemenyi post-hoc test.

Model	Mean (%) ± SD	Average rank	Group (Nemenyi)
LightGBM	90.49 ± 1.38	2.90	A
Random Forest	90.16 ± 1.70	3.45	A
Stacking	90.21 ± 1.70	3.70	A
Extra Trees	89.69 ± 1.55	4.05	A
XGBoost	89.54 ± 1.70	4.45	AC
Bagging	89.59 ± 1.87	5.45	AC
CatBoost	89.31 ± 1.54	5.55	AC

 doi:10.20944/preprints202604.0697.v1

Voting	88.17 ± 2.44	6.85	AB
Gradient Boosting	85.84 ± 3.03	8.60	BC
AdaBoost	75.48 ± 4.20	10.00	B

Note. The mean (%) ± SD corresponds to cross-validation. The average ranks are derived from the Friedman test, and the groups from the Nemenyi post-hoc test (α = 0.05). Models sharing the same letter do not differ significantly.

To ensure complete statistical traceability, Table 3 presents the significant post-hoc comparisons derived from the Nemenyi test. This table enables the explicit identification of pairs of models that exhibit statistically significant differences. It also quantifies the magnitude of these differences in terms of rank and precision, thereby complementing the overall summary presented in Table 2.

Table 3. Significant post-hoc comparisons (Nemenyi, α = 0.05).

Best (rank)	**Worst (rank)**	**p (Nemenyi)**	**Δ rank**	**Δ precision (pp)**	**Conclusion**
LightGBM (2.90)	AdaBoost (10.00)	6.99×10^{-6}	7.10	15.02	Best > Worst*
Random Forest (3.45)	AdaBoost (10.00)	5.75×10^{-5}	6.55	14.69	Best > Worst*
Stacking (3.70)	AdaBoost (10.00)	1.42×10^{-4}	6.30	14.73	Best > Worst*
Extra Trees (4.05)	AdaBoost (10.00)	4.71×10^{-4}	5.95	14.21	Best > Worst*
LightGBM (2.90)	Gradient Boosting (8.60)	0.001062	5.70	4.66	Best > Worst*
XGBoost (4.45)	AdaBoost (10.00)	0.0017	5.55	14.07	Best > Worst*
Random Forest (3.45)	Gradient Boosting (8.60)	0.005554	5.15	4.32	Best > Worst*
Stacking (3.70)	Gradient Boosting (8.60)	0.01104	4.90	4.37	Best > Worst*
Bagging (5.45)	AdaBoost (10.00)	0.02692	4.55	14.11	Best > Worst*
Extra Trees (4.05)	Gradient Boosting (8.60)	0.02692	4.55	3.85	Best > Worst*
CatBoost (5.55)	AdaBoost (10.00)	0.03417	4.45	13.83	Best > Worst*

*Significant difference according to Nemenyi post-hoc test (α = 0.05). Δ precision in pp; ranks are from the Friedman test (lower = better).

Tables 2 and 3 offer a synopsis of the comparative evaluation of the performance of ensemble learning models in predicting obesity. As illustrated in Table 2, the mean accuracy (mean (%) ± SD) was obtained via stratified cross-validation. The average rank, derived from the Friedman test (where lower values indicate better performance), is also reported. The homogeneity groups were defined based on the post-hoc contrast of the Nemenyi post-hoc test (α = 0.05). This analytical framework facilitates the coherent integration of absolute performance, stability across partitions, and inferential evidence under a controlled multiple comparison scheme.

 doi:10.20944/preprints202604.0697.v1

It is important to acknowledge that the minor discrepancies observed between the average metrics reported in the cross-validation and those considered in the Friedman–Nemenyi analysis are attributable to variations in the aggregation procedure. Specifically, the nonparametric test is predicated on per-fold rankings and a strictly aligned evaluation matrix across models, which may introduce marginal variations in the aggregated metrics. However, these differences are negligible and do not compromise the validity or consistency of the inferential conclusions derived from the statistical analysis.

The leading group (Group A) consists of LightGBM (90.49% ± 1.38; rank = 2.90), Random Forest (90.16% ± 1.70; rank = 3.45), Stacking (90.21% ± 1.70; rank = 3.70), and Extra Trees (89.69% ± 1.55; rank = 4.05). Despite LightGBM demonstrating the highest mean accuracy and the lowest average rank, the differences between these models are not statistically significant according to the Nemenyi post-hoc test, suggesting equivalent performance from an inferential standpoint.

At a second level, XGBoost (89.54% ± 1.70; rank = 4.45), Bagging (89.59% ± 1.87; rank = 5.45), and CatBoost (89.31% ± 1.54; rank = 5.55) fall into the AC group, indicating a simultaneous statistical overlap with the top block and with intermediate performance levels. Meanwhile, Voting (88.17% ± 2.44; rank = 6.85), classified in the AB group, shows no significant differences either compared to the leading models or compared to those with moderate performance. This pattern suggests the presence of an inferential transition zone, in which differences in mean accuracy (≈1–2 pp relative to block A) do not translate into statistically significant contrasts.

A decline in performance is observed at the bottom of the ranking, and it is more consistent. Gradient Boosting (85.84% ± 3.03; rank = 8.60), which belongs to group BC, exhibits a significant performance deficit relative to the leading group, although it maintains a certain degree of overlap with intermediate models. Conversely, AdaBoost (75.48% ± 4.20; rank = 10.00), classified exclusively in group B, was identified as the worst-performing model, exhibiting no evidence of statistical equivalence to the most competitive methods.

As illustrated in Table 3, the statistically significant differences identified through the Nemenyi post-hoc test ($\alpha = 0.05$) are predominantly concentrated between the highest-performing models and those at the lower end of the ranking, thereby forming a clearly stratified structure.

Specifically, the models from the leading group—LightGBM (rank = 2.90), Random Forest (3.45), Stacking (3.70), and Extra Trees (4.05)—exhibited consistent superiority over AdaBoost (10.00), with highly significant p-values (6.99×10^{-6} to 4.71×10^{-4}) and substantial disparities in both rank ($\Delta \approx 5.95$–7.10) and precision (≈14.21–15.02 pp). These figures not only substantiate statistical superiority but also reflect a salient practical advantage, thereby reinforcing the congruence between inferential evidence and descriptive performance.

In a similar vein, LightGBM, Random Forest, Stacking, and Extra Trees exhibited a notable enhancement in performance, surpassing Gradient Boosting (rank = 8.60), albeit with comparatively modest margins (Δ precision ≈3.85–4.66 pp; Δ rank ≈4.55–5.70). This pattern suggests that, while Gradient Boosting does not reach the level of the leading model, its performance falls in an intermediate rank, distant from the most competitive models but not as far behind as AdaBoost.

In contrast, intermediate models such as XGBoost (4.45), Bagging (5.45), and CatBoost (5.55) also demonstrate substantial disparities when compared to AdaBoost ($p < 0.05$), with accuracy discrepancies surpassing 13.8 pp. This further substantiates the finding that AdaBoost exhibits suboptimal structural performance.

The incorporation of explainability techniques, such as SHAP and LIME, complemented by calibration analyses, enhances the interpretability and transferability of the results in applied contexts. Specifically, the global importance and partial dependencies derived from SHAP, in conjunction with the local explanations provided by LIME, enable the prediction to be disaggregated into contributions attributable to each variable, thereby facilitating model traceability. This approach enhances the transparency of the predictive system and enables auditing processes aimed at evaluating the stability, consistency, and clinical plausibility of the generated decisions.

 doi:10.20944/preprints202604.0697.v1

As shown in **Figure 1**, the SHAP beeswarm plot for the Extra Trees model in a one-vs-rest configuration (Class 4) demonstrates that the predictive signal is predominantly influenced by a limited set of variables that exhibit substantial clinical plausibility. The three most influential predictors are frequency_of_consuming_vegetables, number_of_main_meals_daily, and age, followed by smoking status (smoking_yes, smoking_no), type of transportation used (type_of_transportation_used_yes/no), and family history of overweight/obesity (overweight_obese_family_yes/no), while calculation_of_calorie_intake_no also makes a significant contribution to the log-odds ratios for belonging to Class 4.

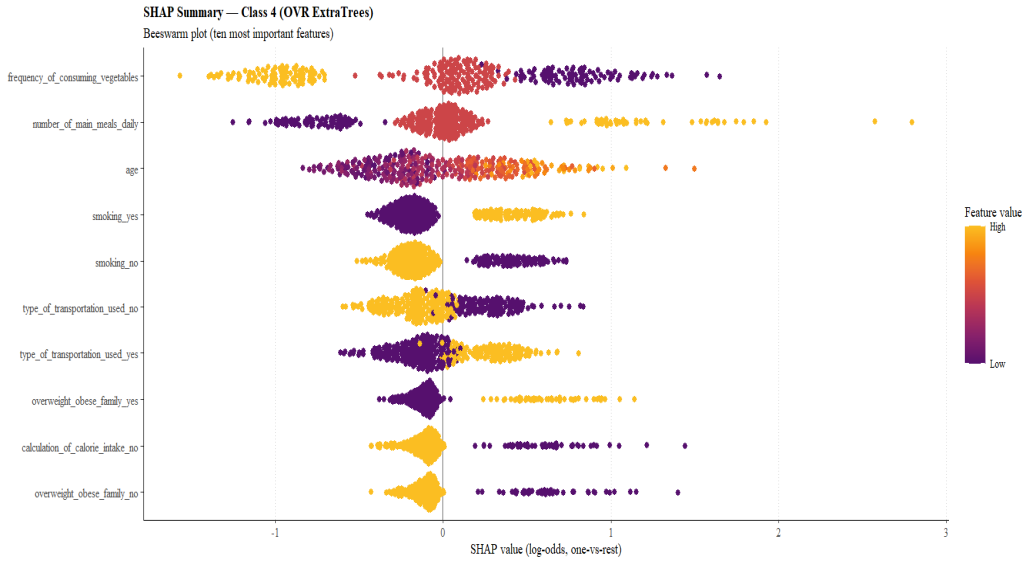

Figure 1. HAP beeswarm for **class 4** (one-vs-rest multiclass scheme using Extra Trees). **Note.** Each point represents an observation from the test set. The horizontal position indicates the **SHAP value** on a **log-odds scale** for the binary classifier for class 4; the colors encode the **feature value** (low→high). The variables are sorted by their **overall importance** in this class (mean of |SHAP|); points to the right **increase** the likelihood of class 4, and those to the left **decrease** it. The apparent discontinuities result from **one-hot encoding**/standardization and **do not** constitute clinical thresholds.

As further illustrated in Figure 1, frequency_of_consuming_vegetables exhibits a clear color gradient: low values (purple) are concentrated on the right side with positive SHAP values, while high values (yellow) are concentrated on the left with negative values. This pattern suggests a potential causal relationship between vegetable consumption and Class 4 classification. Specifically, low vegetable consumption appears to be associated with an increased likelihood of being classified into Class 4, while higher vegetable consumption seems to offer a protective effect. A similar, though less pronounced, gradient is observed for age, where older ages are associated with positive SHAP contributions, consistent with a higher probability of belonging to Class 4 in older subjects.

The complementary behavior of the smoking_yes and smoking_no variables confirms the model's internal consistency. Specifically, active smoking tends to shift predictions toward Class 4 (positive SHAP values for smoking_yes and negative values for smoking_no), while non-smokers are associated with predominantly negative contributions. The variables encoding mode of transportation and family history (type_of_transportation_used_yes/no, overweight_obese_family_yes/no) demonstrate considerable dispersion of SHAP values, suggesting heterogeneous yet directionally consistent effects, in which more sedentary transportation patterns and the presence of family history generally increase the predicted risk.

Finally, the calculation_of_calorie_intake_no model consistently yields predominantly positive SHAP values, suggesting that the failure to monitor caloric intake is associated with an elevated probability of belonging to Class 4. These inferences are strictly local and predictive. The SHAP

 doi:10.20944/preprints202604.0697.v1

values are expressed in log-odds, and therefore, the resulting probability is obtained by applying the logistic function to the sum of the base value and the contributions. The observable discontinuities are attributable to standardization and one-hot encoding, and as such, they should not be interpreted as clinical thresholds. It is imperative to interpret the outcomes within the framework of the binary Class 4 classifier and the one-vs-rest paradigm. This interpretation should not be construed as implying causality. To obtain a more comprehensive perspective, it is recommended to compare the outcomes with global measures of absolute mean importance and to conduct additional out-of-sample validation.

As shown in **Figure 2**, the SHAP waterfall plot corresponds to a single observation classified into Class 4 under a one-vs-rest scheme using Extra Trees. The horizontal axis is plotted on a log-odds scale. The model commences from the baseline value E[f(x)]; this represents the classifier's expected value for that class in the training set. Each bar adds a local contribution, termed "ΔSHAP," associated with a specific feature. The probability of a bar shifting to the right increases the log-odds of belonging to Class 4, while a shift to the left would decrease them. The cumulative sum of all contributions yields the final value f(x), whose probability is obtained via the logistic transformation $\sigma(f) = 1/(1 + \exp(f))$ (in this case, $f(x) \approx 3.94$, which translates to $p \approx 0.978$).

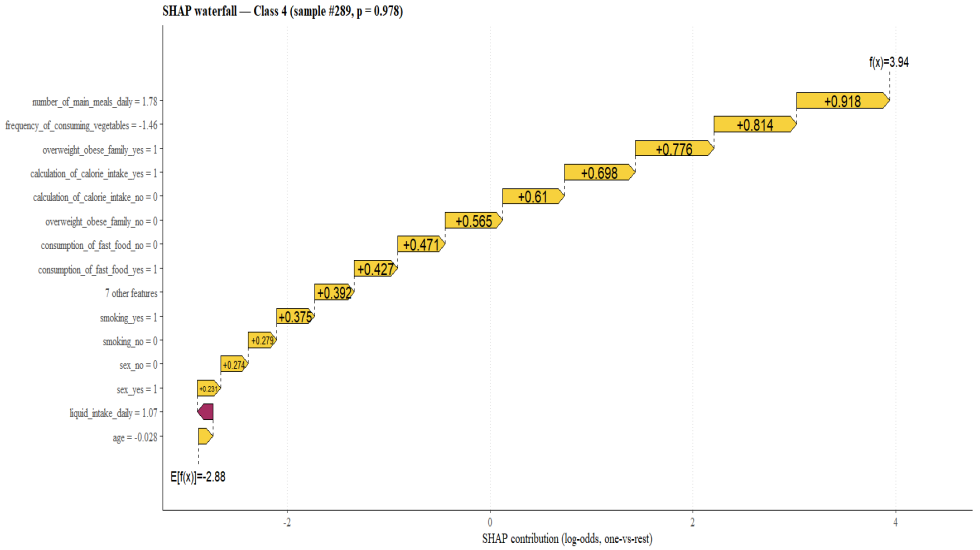

Figure 2. Local SHAP contributions (waterfall) for the predicted class in a one-vs-rest multiclass setting using Extra Trees. **Note.** The plot shows the local breakdown of the prediction for an individual instance. It starts with the expected value E[f(x)] (base), and each bar shows the SHAP contribution ΔSHAP of a feature on the **log-odds** scale of the binary classifier for the **positive class**: bars to the right increase the likelihood; those to the left decrease it. The cumulative sum yields f(x) and the probability $p=\sigma(f)$. Categorical variables are one-hot encoded; apparent cutoff points are not clinical thresholds.

In this instance, the trend is predominantly positive, reflecting substantial evidence in favor of Class 4. The binary or one-hot encoded predictors that are present in the initial rows—such as the number of daily main meals, the frequency of vegetable consumption, the presence of a family history of overweight/obesity, and the calculation of caloric intake—act as the primary drivers of the increase in log-odds, with each yellow bar quantifying the marginal contribution of the corresponding characteristic. Shorter bars are indicative of more modest local effects. This interpretation is strictly local and not causal; it describes how the specific combination of values in this observation modifies the model's output and should not be extrapolated to global rules or clinical cutoffs. To achieve a comprehensive global characterization of the importance of the variables, it is necessary to supplement this analysis with aggregated summaries, such as the mean |SHAP| per feature. If

 doi:10.20944/preprints202604.0697.v1

deemed necessary, resampling procedures (e.g., bootstrapping) or training repetitions should be employed to assess the stability of these contributions.

As shown in Figure 3, LIME illustrates the local interpretability of the multiclass prediction by identifying the features that contribute most to the model's decision. As illustrated in Panel (a), the classifier assigns the analyzed instance to Class 3 with a calibrated probability of 99.4%, in contrast to an uncalibrated probability of 75.3%. This discrepancy suggests that the base model exhibited a tendency toward underconfidence in the high-certainty region. Multinomial calibration rectifies this deviation, yielding probability scores that are consistent with the observed frequency. The remaining classes exhibit residual probabilities that approach zero, a result that positions the instance at a considerable distance from the decision boundary.

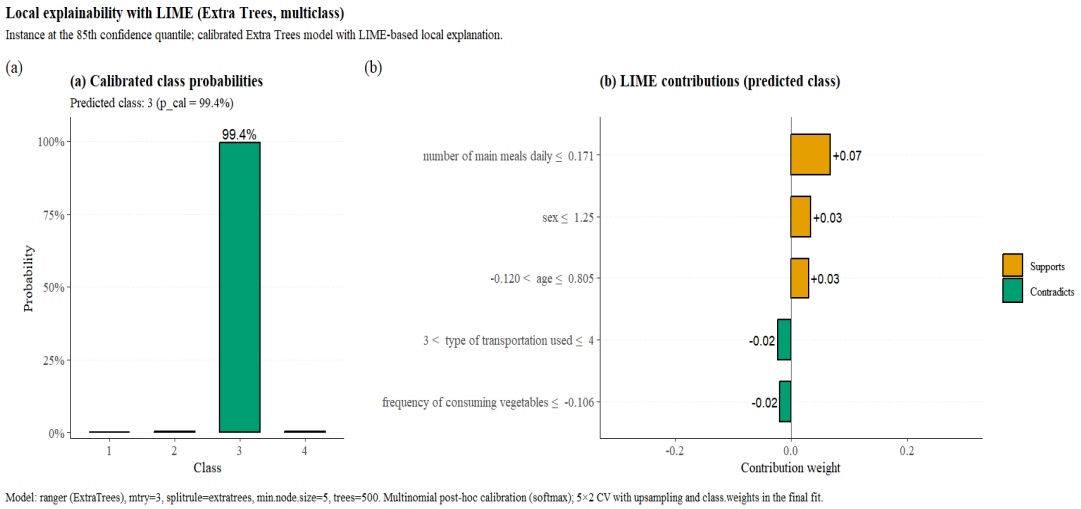

Figure 3. Local explainability with LIME in a multiclass classifier. **Note:** Panel (a) shows calibrated probabilities by class, demonstrating an improvement in the probabilistic consistency of the predictions. Panel (b) presents local explanations that break down the contribution of variables to individual predictions, highlighting that the observed cutoff points correspond to technical discretizations rather than clinical thresholds.

Panel (b) illustrates the local contributions of LIME for the predicted class. The effects are of moderate magnitude and, for the most part, favor Class 3, a pattern that is expected when model confidence is high. The transformations of the variables "number of main meals daily," "sex," and "age" stand out as terms with the highest positive weights, while "type of transportation used" and "frequency of consuming vegetables" exhibit small-scale negative contributions. It is imperative to underscore that the observed inequalities and cutoff points are a consequence of the discretization implemented by LIME on standardized predictors. These should not be interpreted as clinical thresholds. Additionally, the signs and magnitudes are indicative of local relationships within the context of the observed data and do not represent global causal effects of the model.

Indeed, the graphical evidence supports the conclusion that the model, after calibration, produces well-fitted probabilities and that the Class 3 prediction for the analyzed instance is robust. LIME's local decomposition indicates that the decision is the result of the aggregate of numerous small, yet consistent, effects rather than a single dominant predictor. This finding aligns with the nonlinear characteristics of the classifier. For substantive interpretation, it is important to note that the results are specific to the instance and the neighborhood considered by LIME.

5. Discussion

The findings do not provide substantial evidence to support the hypothesis that a single model dominates the top-performing group. Instead, they suggest the presence of a set of leading models that are statistically indistinguishable. In this context, Stacking emerges as a highly competitive

 doi:10.20944/preprints202604.0697.v1

method (average rank = 3.70), falling into Group A in Nemenyi post-hoc comparison. However, LightGBM (2.90), Random Forest (3.45), and Extra Trees (4.05) were found to be statistically indistinguishable, suggesting that no substantial differences exist among these models after adjustment for multiple comparisons.

From a methodological perspective, this convergence between ranking and overall performance suggests that leadership should be interpreted in terms of inferential equivalence rather than absolute superiority. In particular, the utilization of average ranks derived from the Friedman test is deemed appropriate in contexts where performance may vary across partitions due to data heterogeneity. Under this approach, Stacking maintains competitive and stable performance, consistent with the logic of variance-reduction-oriented meta-ensembles. However, it does not demonstrate a statistically significant advantage over other models in the top tier.

Utilizing XGBoost, LightGBM, and CatBoost—which have gained renown for their efficacy with tabular data—as a foundation, the findings substantiate that these models offer a reliable starting point. However, their placement in the AC groups (XGBoost and CatBoost) or even within Group A itself (LightGBM) reveals that, while competitive, they do not demonstrate consistent superiority over the set of leading models. In particular, LightGBM achieves the best overall average rank; however, there are no significant differences in performance compared to Random Forest, Stacking, and Extra Trees, which reinforces the idea of statistical equivalence in the top tier.

In summary, rather than identifying a single "best model," the empirical evidence suggests the existence of a range of statistically comparable high-performance models, within which Stacking is a robust—but not exclusive—alternative. This finding carries substantial practical implications, as it facilitates model selection based on supplementary criteria, such as computational cost, interpretability, or ease of implementation, without substantial compromise to predictive accuracy.

This choice is consistent with empirical and applied evidence in the literature. In the early prediction of childhood obesity using EHR data, XGBoost can achieve outstanding performance (AUC ≈ 0.81), demonstrating why it is used as a strong benchmark [42]. In a similar vein, LightGBM has been validated in clinical cohorts due to its capacity to capture nonlinear interactions and its natural integration with interpretability frameworks such as TreeSHAP [39], thereby reinforcing its role as a robust baseline in biomedical settings.

The results obtained from the post-hoc analysis reveal that, while the baselines remain within Group A, the average ranks are less favorable than those of Stacking. This phenomenon, characterized as "comparable" under the Nemenyi threshold yet exhibiting systematically worse rankings, is to be anticipated when differences are moderate, and variability across partitions impedes the capacity to discern high-performing methods. In practical terms, the conclusion is not that the baselines are inadequate; rather, it is that meta-ensembles tend to capture complementarities that manifest as overall consistency rather than drastic jumps in a single metric. This pattern aligns with recent findings in the field, which indicate that hybrid combination strategies often outperform individual models. For instance, a majority voting approach that integrates Gradient Boosting, XGBoost, and MLP has been shown to improve performance relative to the best individual model [43]. This enhancement in performance can be attributed to the diversity of errors contributed by each component.

The Friedman–Nemenyi test is a statistical method that partitions the comparative space into interpretable strata. Group A encompasses methods with medium-to-high precision, exhibiting no discernible intragroup variations below the post-hoc threshold. This observation should be interpreted as inadequate evidence to substantiate statistical superiority among specific pairs, rather than as an indication of equality. In this context, the average rank provides complementary value by synthesizing the accumulated evidence through partitions. The concept of stacking, which involves maintaining the highest rank, has been demonstrated to concentrate the signal of relative dominance.

Group B, which includes Gradient Boosting and AdaBoost, exhibits substandard performance in terms of inferential separation. This finding aligns with prior observations indicating that classical models or those with reduced robustness tend to demonstrate diminished stability in real-world

 doi:10.20944/preprints202604.0697.v1

clinical contexts. In preliminary studies, the efficacy of tree-based methods (e.g., ID3) in predicting childhood obesity was found to be inferior to the performance of modern ensembles. This observation highlights the significant advancements in robustness and generalization capability that have been achieved through the adoption of contemporary approaches.

In healthcare contexts, the usefulness of a model is not limited to classification; clinical adoption requires interpretability and plausibility. The extant literature, as referenced in the matrix, aligns in unison on this matter. For instance, the study by Fu et al. [39] employs SHAP/TreeSHAP to interpret a LightGBM-based model and translate it into early risk factors, underscoring that explanation is not merely an ancillary feature but rather an integral component of clinical value. Conversely, Helforoush and Sayyad [46] employed SHAP to elucidate a hybrid approach (ANN-PSO) aimed at risk profiling, underscoring how interpretation reinforces confidence and preventive action. In adults, the application of interpretable learning to quantify the contribution of lifestyle factors (e.g., sedentary behavior, alcohol, and protein) has also been documented using SHAP in boosted tree models [47]. A recent narrative review has concluded that, in the context of obesity, ML algorithms (e.g., Random Forest, XGBoost) and the interpretability of models (with SHAP being the predominant tool) are emerging as recurrent methodological pillars [40].

In addition, it is recommended to incorporate calibration measures, such as the Brier score and reliability plots, along with classification metrics. This is due to the fact that, in clinical settings, decision-making is informed by reliable probabilities rather than definitive labels. A review/tutorial framework centered on the identification of risk factors and model performance also explicitly highlights discrimination and calibration as relevant reporting dimensions in obesity/overweight [58]. Consequently, Stacking's leadership in discrimination and ranges should be accompanied by an assessment of calibration to support risk-oriented conclusions.

From an implementation perspective, the advantages of Stacking must be weighed against its higher computational cost. In an offline/batch scenario, the approach is justifiable as a high-performance option; however, if latency or resource constraints are a concern, alternatives with comparable performance (e.g., Random Forest/Extra Trees-type ensembles) may be preferable due to their cost–benefit ratio, while also maintaining reasonable interpretability when XAI tools are integrated. In summary, the results are consistent with strong baselines [40], improvements through model ensembling, and the need to integrate explainability and clinical utility criteria [46], with the methodological addition of calibration as a key reporting element.

6. Conclusions

The findings indicate that ensemble learning methods are remarkably effective strategies for predicting obesity risk in settings characterized by population heterogeneity. However, no single model demonstrates statistical superiority. Instead, a leading group of models that are statistically indistinguishable from one another emerges. This group comprises LightGBM, Random Forest, Stacking, and Extra Trees, all of which belong to Group A in the Nemenyi post-hoc test. Within this set, LightGBM presents the best average rank, although without significant differences compared to the other models in the top block, which reinforces the notion of inferential equivalence in terms of accuracy.

In this context, Stacking emerges as a robust alternative among the highest-performing models, consistent with the principle of meta-ensembles to capture complementarities among predictors and reduce variance. However, it is imperative to interpret its superiority with caution, as there is an absence of evidence substantiating a statistically significant advantage over Random Forest, Extra Trees, or LightGBM. Consequently, the selection of a model should not be predicated exclusively on marginal disparities in accuracy, but rather on a comprehensive analysis that also considers operational criteria.

In contrast, AdaBoost demonstrated substandard performance, exhibiting statistically significant discrepancies compared to numerous top-performing models. Gradient Boosting, conversely, exhibited performance in the lower-middle range, manifesting substantial disparities

 doi:10.20944/preprints202604.0697.v1

compared to the leading models. The findings indicate that neither approach is a preferred option for this task within the evaluated parameters.

From an applied perspective, the existence of a set of statistically equivalent models enables an informed selection based on computational efficiency, interpretability, and ease of deployment, without substantially compromising predictive performance. In this regard, Random Forest and Extra Trees emerge as particularly attractive alternatives due to their balance between performance and computational cost. Conversely, Stacking can be considered a suitable option when maximum predictive power is prioritized in environments where computational cost is not a constraint.

In contemplating future research endeavors, it is imperative to enhance the clinical validity and generalizability of the models through the following measures: The following five points must be considered: (i) systematic calibration assessment (e.g., Brier score and reliability plots); (ii) multicenter external validation across varying prevalence rates and covariates; (iii) equity audits in population subgroups; (iv) incorporation of reproducible explainability techniques (such as SHAP) alongside stability analyses; and (v) prospective monitoring of model drift and degradation in real-world settings. The evidence supports the use of advanced ensembles as a reference strategy, not from the perspective of a single dominant model, but as a set of robust and statistically equivalent solutions. The selection of these solutions must align with the operational conditions and clinical objectives of the decision support system.

Author Contributions: Conceptualization, Daniel Andrade-Girón and William Marin-Rodriguez; Methodology, Daniel Andrade-Girón and William Marin-Rodriguez; Software, Daniel Andrade-Girón; Validation, Daniel Andrade-Girón, Américo Peña, Elsa Oscuvilca-Tapia and Fredy Bermejo-Sanchez; Formal analysis, Daniel Andrade-Girón, William Marin-Rodriguez and Américo Peña; Investigation, Daniel Andrade-Girón, William Marin-Rodriguez and Américo Peña; Resources, Américo Peña, Elsa Oscuvilca-Tapia and Fredy Bermejo-Sanchez; Writing—original draft, Daniel Andrade-Girón and William Marin-Rodriguez; Writing—review & editing, Daniel Andrade-Girón and William Marin-Rodriguez; Visualization, Daniel Andrade-Girón, Elsa Oscuvilca-Tapia and Fredy Bermejo-Sanchez; Supervision, William Marin-Rodriguez; Project administration, Américo Peña, Elsa Oscuvilca-Tapia and Fredy Bermejo-Sanchez; Funding acquisition, Américo Peña, Elsa Oscuvilca-Tapia and Fredy Bermejo-Sanchez.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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). No new data were created. The dataset is accessible to the research community through the Kaggle platform.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

NOTE: The enhancement of the article's quality was achieved through the utilization of artificial intelligence platforms, though not as a standalone mechanism for generating the article itself.

References

1. Mehmood, D.A. Epidemiological Trends and Risk Factors of Non-Communicable Diseases: A Global Perspective. *Noncommun Dis Care* **2025**, *2*, 12–22.
2. Colmenarejo, G. Machine Learning Models to Predict Childhood and Adolescent Obesity: A Review. *Nutrients* **2020**, *12*, 2466, doi:10.3390/nu12082466.

 doi:10.20944/preprints202604.0697.v1

3. Ferdowsy, F.; Rahi, K.S.A.; Jabiullah, Md.I.; Habib, Md.T. A Machine Learning Approach for Obesity Risk Prediction. *Curr Res Behav Sci* **2021**, *2*, 100053, doi:10.1016/j.crbeha.2021.100053.
4. Vekic, J.; Stefanovic, A.; Zeljkovic, A. Obesity and Dyslipidemia: A Review of Current Evidence. *Curr Obes Rep* **2023**, *12*, 207–222, doi:10.1007/s13679-023-00518-z.
5. Pang, H.; Zhou, L.; Dong, Y.; Chen, P.; Gu, D.; Lyu, T.; Zhang, H. Electronic Health Records-Based Data-Driven Diabetes Knowledge Unveiling and Risk Prognosis. *ICCK Trans Intell Syst* **2024**, *2*, 1–13, doi:10.62762/TIS.2025.367320.
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, doi:10.1038/s41569-023-00847-5.
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, doi:10.1016/j.compbiomed.2021.104754.
8. Cheng, E.R.; Steinhardt, R.; Ben Miled, Z. Predicting Childhood Obesity Using Machine Learning: Practical Considerations. *BioMedInformatics* **2022**, *2*, 184–203, doi:10.3390/biomedinformatics2010012.
9. 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, doi:10.1111/obr.12872.
10. Sørensen, T.I.A. Forecasting the Global Obesity Epidemic Through 2050. *The Lancet* **2025**, *405*, 756–757, doi:10.1016/S0140-6736(25)00260-0.
11. Aremu, S.O.; Akute, B.; Aremu, D.O.; Zando, C.; Aremu, E.D.; Nwachukwu, O.J.; Omosebi, 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, doi:10.1186/s12982-025-00818-w.
12. Couto, F. da F.S.; Almeida, C.P.B. de. 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*, doi:10.3390/ijerph22071152.
13. Grossman, E. Obesity and Cardiovascular Disease: Mechanistic Insights and Nutritional Management Strategies. *Nutrients* **2025**, *17*, doi:10.3390/nu17142281.
14. 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, doi:10.1007/s13679-019-00358-w.
15. Qian, F.; Wang, L. The Future of Obesity Management: Bridging Pharmacologic Innovation and Public Health. *Diabetes Metab Syndr Obes* **2025**, *18*, 4667–4670, doi:10.2147/DMSO.S579769.
16. 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, doi:10.1038/s41366-024-01650-z.
17. Ganie, S.M.; Reddy, B.B.; K, H.; Rege, M. An Investigation of Ensemble Learning Techniques for Obesity Risk Prediction Using Lifestyle Data. *Decis Anal J* **2025**, *14*, 100539, doi:10.1016/j.dajour.2024.100539.
18. Airlangga, G. Machine Learning-Based Obesity Classification: A Comparative Study Using Self-Reported Survey Data and Ensemble Learning Models. *J Teknol Inform Komput* **2025**, *11*, 347–361, doi:10.37012/jtik.v11i1.2585.
19. Çizmeci, 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); May 2025; pp. 1–8.
20. 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, doi:10.1109/JBHI.2024.3356580.
21. Barus, O.P.; Billie, F.; Jusin; 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); September 2024; pp. 1–5.

 doi:10.20944/preprints202604.0697.v1

22. 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*, doi:10.3390/ijerph21010002.
23. 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, doi:10.1007/s10994-022-06268-8.
24. Barbierato, E.; Gatti, A. The Challenges of Machine Learning: A Critical Review. *Electronics* **2024**, *13*, 416, doi:10.3390/electronics13020416.
25. Yehia, T.; Wahba, A.; Mostafa, S.; Mahmoud, O. Suitability of Different Machine Learning Outlier Detection Algorithms to Improve Shale Gas Production Data for Effective Decline Curve Analysis. *Energies* **2022**, *15*, doi:10.3390/en15238835.
26. 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.Th., Eds.; Springer International Publishing: Cham, 2022; pp. 79–163 ISBN 978-3-031-00832-0.
27. Xin, R.; Wang, J.; Chen, P.; Zhao, Z. Trustworthy AI-Based Performance Diagnosis Systems for Cloud Applications: A Review. *ACM Comput Surv.* **2025**, *57*, 1-37, doi:10.1145/3701740.
28. Keshavamurthy, D.; Kumar, M.; Tsaramirsis, G.; Oroumchian, F. An AI-Based Framework for Secure and Transparent Banking: Integrating Adversarial Robustness, Interpretability, and Organizational Modeling. *Secur Priv* **2026**, *9*, e70153, doi:10.1002/spy2.70153.
29. 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*, 148, doi:10.1145/3689627.
30. 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. *Arthroscopy* **2025**, *41*, 455–472, doi:10.1016/j.arthro.2024.01.043.
31. 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 Di* **2022**, *14*, e1548, doi:10.1002/wsbm.1548.
32. 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, doi:10.47852/bonviewJDSIS32021149.
33. 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, doi:10.1109/ACCESS.2025.3530840.
34. 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, doi:10.3390/e25020245.
35. 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, doi:10.3390/app152212036.
36. Alsulamy, S. Predicting Construction Delay Risks in Saudi Arabian Projects: A Comparative Analysis of CatBoost, XGBoost, and LGBM. *Exp Syst Appl* **2025**, *268*, 126268, doi:10.1016/j.eswa.2024.126268.
37. Heddam, 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 Switzerland: Cham, 2024; pp. 1–25 ISBN 978-3-031-66842-5.
38. 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, doi:10.4338/ACI-2015-03-RA-0036.
39. 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, doi:10.1016/j.jinf.2020.03.041.

 doi:10.20944/preprints202604.0697.v1

40. Shi, X.; Nikolic, G.; Epelde, G.; Arrúe, M.; Bidaurrazaga 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, doi:10.1186/s12911-021-01580-0.
41. Pang, G.; Shen, C.; Cao, L.; Hengel, A.V.D. Deep Learning for Anomaly Detection: A Review. *ACM Comput Surv* **2021**, *54*, 1–38, doi:10.1145/3439950.
42. 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, doi:10.1093/nar/gkab1107.
43. Solomon, D.D.; Khan, S.; Garg, S.; Gupta, G.; Almjally, A.; Alabduallah, 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, doi:10.3390/diagnostics13152610.
44. Abnoosian, K.; Farnoosh, R.; Behzadi, M.H. Prediction of Diabetes Disease Using an Ensemble of Machine Learning Multi-Classifier Models. *BMC Bioinformatics* **2023**, *24*, 337, doi:10.1186/s12859-023-05465-z.
45. 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, doi:10.1089/jpm.2023.0087.
46. Helforoush, Z.; Sayyad, H. Prediction and Classification of Obesity Risk Based on a Hybrid Metaheuristic Machine Learning Approach. *Front Big Data* **2024**, *7*, doi:10.3389/fdata.2024.1469981.
47. 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, doi:10.1186/s12889-024-20510-z.
48. Taherkhani, A.; Cosma, G.; McGinnity, T.M. AdaBoost-CNN: An Adaptive Boosting Algorithm for Convolutional Neural Networks to Classify Multi-Class Imbalanced Datasets Using Transfer Learning. *Neurocomputing* **2020**, *404*, 351–366, doi:10.1016/j.neucom.2020.03.064.
49. Zhou, Y.; Mazzuchi, T.A.; Sarkani, S. M-AdaBoost-A Based Ensemble System for Network Intrusion Detection. *Exp Syst Appl* **2020**, *162*, 113864, doi:10.1016/j.eswa.2020.113864.
50. 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 Intel* **2020**, *94*, 103770, doi:10.1016/j.engappai.2020.103770.
51. Konstantinov, A.V.; Utkin, L.V. Interpretable Machine Learning with an Ensemble of Gradient Boosting Machines. *Knowl Based Syst* **2021**, *222*, 106993, doi:10.1016/j.knosys.2021.106993.
52. 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); December 2024; pp. 667–672.
53. Dagnew, G.; Shekar, B. H. Ensemble Learning-Based Classification of Microarray Cancer Data on Tree-Based Features. *Cogn Comput Syst* **2021**, *3*, 48–60, doi:10.1049/ccs2.12003.
54. 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 Prob Eng* **2020**, *2020*, 2835023, doi:10.1155/2020/2835023.
55. Trujillano, J.; Serviá, L.; Badia, M.; Serrano, J.C.E.; Bordejé-Laguna, M.L.; Lorencio, C.; Vaquerizo, C.; Flordelis-Lasierra, J.L.; Lagrán, I.M. de; Portugal-Rodríguez, E.; et al. Methodological Review of Classification Trees for Risk Stratification: An Application Example in the Obesity Paradox. *Nutrients* **2025**, *17*, 1903, doi:10.3390/nu17111903.
56. Nitha, V.R.; Vinod Chandra, S.S. ExtRanFS: An Automated Lung Cancer Malignancy Detection System Using Extremely Randomized Feature Selector. *Diagnostics* **2023**, *13*, 2206, doi:10.3390/diagnostics13132206.
57. Shaban, W.M.; El-Din Moustafa, H.; El-Seddek, M.M. Machine Learning Framework for Predicting Susceptibility to Obesity. *Sci Rep* **2025**, *15*, 35040, doi:10.1038/s41598-025-20505-9.
58. 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, doi:10.3390/s20092734.

 doi:10.20944/preprints202604.0697.v1