Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 2, 2026

EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) FOR CLASSIFICATION OF OBESITY: A COMPARATIVE STUDY BASED ON LOCAL AND GLOBAL INTERPRETABILITY

[1]Adam Khan, [*2]Jahangir Khan
[1]Department of Computer Science and IT, Sarhad University of Science and Information Technology, Peshawar 25000, Khyber Pakhtunkhwa, Pakistan. adam.me@suit.edu.pk
*[*2]Department of Computer Science and IT, Sarhad University of Science and Information Technology, Peshawar 25000, Khyber Pakhtunkhwa, Pakistan. Corresponding Author Email: jahangir.csit@suit.edu.pk*

DOI: https://doi.org/10.5281/zenodo.24108203

Keywords
Explainable artificial intelligence; obesity classification; LIME; permutation feature importance; random forest; logistic regression; interpretability

Article History
Received on 14 Jan, 2026
Accepted on 10 Feb, 2026
Published on 18 Feb, 2026

Copyright @Author
Corresponding Author: Jahangir Khan

Abstract

Obesity is one of the leading global health problems that cause the development of chronic diseases, thus the development of highly accurate and interpretable predictive models is necessary for prevention. While there have been numerous successful applications of machine learning (ML) to classify cases of obesity, most of them focus on high accuracy rather than interpretability and hardly evaluate agreement between different explanation techniques. The current paper proposes an interpretable ML approach for seven-class obesity classification based on the lifestyle dataset with 2,111 samples. Logistic Regression (LR), Random Forest (RF), k-Nearest Neighbors (KNN), and Decision Tree (DT) were developed using the caret package in R. Stratified 80/20 train-test splitting and five-fold cross-validation were used. Accuracy, macro F1-score, Cohen's Kappa, and area under the ROC curve (AUC) were used as evaluation metrics. Permutation Feature Importance (PFI) and Local Interpretable Model-Agnostic Explanations (LIME) were used to derive global and local explanations respectively. LR and RF achieved the highest performance (Accuracy = 96.19%, Kappa = 0.956), RF showed the best discrimination (AUC = 0.9988), KNN reached an accuracy of 0.812, and DT was the weakest (0.545) because its tuned structure relied almost exclusively on Weight and Gender. Weight was the dominant predictor under PFI for all four models and under LIME for DT, LR, and RF, whereas Height was important mainly for LR and RF. LIME explanations for KNN were led by behavioral variables (SMOKE, CAEC, and SCC). Agreement between LIME and PFI was strong for RF (r = 0.969, ρ = 0.817), moderate for LR (r = 0.854, ρ = 0.564), artificially inflated for DT (r = 0.999, ρ = 0.592), and absent for KNN (r = −0.106, ρ = 0.097). Overall, RF offered the best balance of accuracy, robustness, and explanation consistency, while LR provided the most clinically transparent structure. Because roughly 77% of the records are synthetic, the findings should be read as a methodological comparison rather than as evidence of real-world performance.

1. Introduction

Obesity has emerged as a critical global health challenge, contributing significantly to the rising burden of chronic conditions such as cardiovascular disease, type 2 diabetes, and metabolic disorders [1]. According to the World Health Organization (WHO), obesity prevalence has increased substantially over recent decades, making it one of the leading causes of preventable mortality worldwide [2]. The complex and multifactorial nature of obesity, driven by interactions among physiological, behavioral, and environmental factors, poses significant challenges for accurate prediction, risk stratification, and early intervention [3], [4].

Traditional statistical approaches are often limited in their ability to model such complexity because they rely on linear assumptions and have a restricted capacity to capture high-dimensional and nonlinear

relationships [5]. In contrast, supervised machine learning (ML) models have shown strong potential in obesity prediction by learning complex patterns from heterogeneous data sources, including demographic attributes, dietary habits, and lifestyle behaviors. These approaches have been shown to improve predictive performance and to support data-driven healthcare decision-making [6]–[8].

Despite these advances, a critical barrier to the adoption of ML in healthcare is the lack of interpretability. Many high-performing models, particularly ensemble and deep learning approaches, function as black boxes and provide limited insight into the underlying decision-making process. This lack of transparency reduces trust among clinicians and limits the practical deployment of such models in real-world healthcare systems [9]–[12]. Consequently, Explainable Artificial Intelligence (XAI) has gained increasing attention as a means of improving model transparency and interpretability [13], [14]. Techniques such as Local Interpretable Model-Agnostic Explanations (LIME) [15] enable instance-level understanding of predictions by identifying the feature contributions behind individual cases.

However, existing research on obesity prediction presents several limitations. First, many studies prioritize predictive accuracy without systematically evaluating the interpretability of the models or the reliability of their feature explanations [16]. Second, although XAI techniques such as LIME are frequently applied, their consistency with global feature-importance methods is rarely examined, which raises concerns about the stability and trustworthiness of explanations across models and instances. Third, prior work often focuses on complex hybrid or deep learning architectures [17], overlooking the potential of classical ML models to achieve comparable performance with better interpretability. Finally, the reliance on partially synthetic datasets introduces additional challenges for assessing real-world generalizability [18]–[20].

To address these limitations, this study proposes a comprehensive and interpretable ML framework for multi-class obesity classification. Unlike approaches that focus solely on predictive performance, this work emphasizes both model accuracy and interpretability reliability. Specifically, the study integrates multiple classical ML models within a unified experimental framework and evaluates them from both predictive and explainability perspectives. The key contributions are as follows:

- A unified comparative analysis of four supervised ML models, namely Logistic Regression [21], Random Forest [22], k-Nearest Neighbors [23], [24], and Decision Tree [25], [26], under a consistent experimental setting.
- A dual-level interpretability framework that combines global feature importance (Permutation Feature Importance) [27], [28] with local explanations (LIME) [15].
- A correlation-based analysis (Pearson and Spearman) that evaluates the consistency between global and local explanations [15], [27].
- A systematic assessment of feature stability and model reliance [27] across different algorithms.
- Evidence that simpler models can achieve competitive performance while offering more consistent explanations.

By combining predictive performance with multi-level interpretability analysis, this study aims to provide reliable, transparent, and clinically meaningful models for obesity classification that support more informed and trustworthy healthcare decision-making.

The remainder of the paper is organized as follows. Section 2 reviews related work, Section 3 describes the dataset, models, and XAI methodology, Section 4 reports the predictive and explainability results, Section 5 discusses the findings, Section 6 outlines the limitations, and Sections 7 and 8 present the conclusions and directions for future work. As a result, there is an increasing emphasis on applying Explainable Artificial Intelligence (XAI) to increase model transparency and interpretability [13], [14]. Methods such as Local Interpretable Model-Agnostic Explanations (LIME) [15] help in understanding the model prediction at the instance level by identifying which features were significant in making that particular decision.

Nonetheless, there exist certain drawbacks in current literature regarding predicting obesity as well. First of all, many papers have an approach based only on improving the prediction accuracy without checking whether models used are interpretable and reliable in terms of explaining features. Second, even though techniques like LIME are applied widely, consistency with global feature importance approaches is not taken into consideration, thus posing some questions as regards the robustness and reliability of explanations. Third, previous research is mostly centered on using hybrid and/or deep learning frameworks [17] while paying no attention to traditional machine learning algorithms capable of giving comparable prediction accuracy with greater explainability. Last but not least, application of partial synthetically generated data adds another layer of problems [18]–[20].

In order to overcome these shortcomings, this research suggests an interpretable ML framework for multi-class obesity prediction task. As opposed to the approaches which consider the predictive performance alone, this research takes into account both the predictive accuracy and the reliability of interpretation. This paper includes multiple classic machine learning algorithms in one common experimental setup and analyzes their performance both from the perspective of predictive performance and explainability. The main contributions of this paper are the following ones:

- A comparative analysis of four supervised machine learning algorithms, i.e., Logistic Regression [21], Random Forest [22], k-Nearest Neighbors [23], [24] and Decision Tree [25], [26], in a unified experimental setup.
- Dual-level interpretable framework which includes global (Permutation Feature Importance) [27], [28] and local explanations (LIME) [15].
- Correlation analysis based on Pearson and Spearman correlations in order to check the consistence of global and local explanation approaches [15], [27].
- An assessment of feature stability and model dependence [27] for multiple machine learning algorithms.
- The evidence of better performance of simpler algorithms in terms of consistency of explanations..

2. Literature Review

The use of ML for obesity prediction has grown with the increasing availability of structured health and lifestyle datasets. A key resource in this domain is the dataset introduced by Palechor and de la Hoz Manotas, which contains demographic, dietary, and physical-activity attributes for multi-class obesity classification [19]. The dataset has since been used as a benchmark for ML-based obesity prediction and explanation [29].

A substantial body of research indicates that ML models can outperform traditional statistical approaches because they capture nonlinear relationships among complex health variables [30]. Classical algorithms such as Support Vector Machines (SVM) [31], Decision Trees [25], and Random Forests (RF) [22] have been widely applied, with RF showing strong robustness and generalization on health-related data. Deep neural networks [32] extend these capabilities further, but usually at the expense of transparency.

Recent obesity-specific studies (2024–2026) have advanced the field by incorporating optimization, explainability, and real-world datasets. Biswas et al. [33] compared seven ML classifiers on a nationally representative survey dataset and reported that an SVM reached 95.79% accuracy when combined with SMOTE-ENN balancing and hyperparameter tuning, highlighting the importance of data balancing and optimization; they also applied permutation importance and SHAP, but only to the best-performing model. Gandhi et al. [34] benchmarked six ML algorithms and found that RF and Gradient Boosting consistently outperformed the other models, particularly when class imbalance was addressed with SMOTE. The authors Biswas et al. [33] conducted an experiment to compare seven different ML classifiers using a national survey dataset and revealed that an

SVM with 95.79% accuracy was obtained using the combination of SMOTE-ENN technique for balancing and hyperparameter optimization; moreover, Biswas et al. used the techniques of permutation importance and SHAP for evaluation, but not for all of the classifiers. In turn, Gandhi et al. [34] used a benchmarking approach to compare six different ML algorithms and stated that RF and Gradient Boosting are the best performers among other classifiers, especially in cases where class imbalance was balanced using SMOTE.

Some authors have proposed hybrid and optimization-based frameworks as well. Thus, Shafi et al. [35] proposed a hybrid framework of bio-inspired feature selection and hyperparameter optimization; as a result, it turned out that joint optimization allows achieving better results compared to conventional approaches. Moreover, Shaban et al. [36] suggested an ML classifier that incorporates demographic and lifestyle characteristics of people and proved its superiority in obesity level classification.

The use of deep learning techniques has also been investigated. Zeeshan et al. [37] made use of deep neural networks for predicting obesity from dietary surveys and found that they perform better than conventional algorithms like logistic regression and decision tree methods in terms of capturing complex dietary patterns. Alsulami et al. [17] developed an interpretable deep learning model by integrating LSTM with LIME and were able to achieve multi-class obesity prediction with high accuracy while preserving interpretability through visualization and explanations.

XAI has proved to be a very important part of current obesity prediction models. Khater et al. [38] found that ML models with explanations can achieve high predictive performance (up to 86.5% accuracy) without using BMI and using other lifestyle factors instead, thus making them clinically relevant. Görmez et al. [39] combined ML models with SHAP and LIME to predict obesity levels and found that age, weight, height, and certain food patterns were the most important predictors. They found that LIME had higher fidelity while SHAP provided sparser and more consistent explanations among different models. Azad et al. [29] paired a stacking ensemble with LIME on the same UCI dataset used here and reported a peak accuracy of 98.82%. Together, these studies show that explainability supports trust and helps identify key determinants such as physical activity, dietary habits, and anthropometric measures.

Research has been geared towards clinical application as well. For instance, Teke et al. [40] found that ML models can predict treatment compliance and long-term weight loss among obese patients, with an accuracy level of more than 80%.

However, some issues continue to prevail. First, while various studies report high levels of predictive accuracy, the consistency of local interpretability methods (such as LIME) and global methods (such as permutation feature importance) is not analyzed. Secondly, the increasing use of complicated ML models such as deep learning and hybrid optimization systems reduces interpretability and computational efficiency. Thirdly, many studies depend on artificial balancing methods or small samples, which may cause bias and reduced generalizability [18]. Lastly, it is hard to find a framework that would consider several aspects at once, including predictive accuracy, feature importance, and interpretability consistency of different models. Thus, it is needed to develop a standardized framework for ML model assessment that will combine global and local explanations and perform comparison of several ML models. This paper attempts to fill this research gap.

Table 1 summarizes key existing studies, highlighting the models used, the datasets, the main contributions, and the limitations relevant to this work.

Table 1. *Summary of key existing studies on ML-based obesity prediction and explainability.*

Study	Year	Model(s) used	Dataset	Key contribution	Limitation
Breiman [22]	2001	Random Forest	General datasets	Robust ensemble learning with strong generalization	Limited interpretability
Ribeiro et al. [15]	2016	LIME (XAI)	General ML models	Introduced a model-agnostic local interpretability method	Local explanations only; no global consistency check
Palechor & de la Hoz Manotas [19]	2019	Decision Tree, ML models	UCI/Kaggle obesity dataset	Introduced a benchmark multi-class obesity dataset based on lifestyle and physical condition	Contains synthetic data (~77%), with potential bias
Khater et al. [38]	2024	Explainable ML models	Lifestyle datasets	Predicts obesity without BMI (up to 86.5% accuracy); improves clinical relevance	Moderate accuracy compared with ensemble methods
Azad et al. [29]	2025	Stacking ensemble + LIME	UCI obesity dataset (Palechor)	Peak accuracy of 98.82% with LIME explanations	Local explanations only; no global–local consistency analysis
Gandhi et al. [34]	2025	RF, Gradient Boosting, SVM	Obesity datasets	Comparative benchmarking showing RF and boosting outperform other models	No XAI or interpretability analysis
Shaban et al. [36]	2025	ML classifiers	Real-world lifestyle dataset	Integrates demographic and behavioral factors for improved obesity prediction	Limited comparison across models
Zeeshan et al. [37]	2025	Deep Neural Networks (DNN)	Dietary datasets	Shows DNN superiority in capturing complex dietary patterns	Black-box nature (low interpretability)
Görmez et al. [39]	2025	XAI-integrated ML (SHAP, LIME)	Physical activity and eating-habit data	Interpretable insights into obesity risk factors; compares SHAP and LIME	Limited scalability evaluation
Teke et al. [40]	2025	ML models (various)	Clinical datasets	Demonstrates real-world applicability (treatment prediction, >80% accuracy)	General review; lacks experimental validation
Alsulami et al. [17]	2025	LSTM + LIME	Obesity dataset	Combines deep learning with explainability using LIME	Relies on local explanations only
Biswas et al. [33]	2026	SVM, LR, RF, KNN, XGBoost, CatBoost, MLP	National survey dataset (BDHS 2017–18)	95.79% accuracy (SVM) with SMOTE-ENN and tuning; PFI and SHAP on the best model	Binary target; explanations limited to one model

Study	Year	Model(s) used	Dataset	Key contribution	Limitation
Shafi et al. [35]	2026	Hybrid ML + optimization	Health datasets	Bio-inspired feature selection with tuning improves accuracy and robustness	High computational complexity
This study	**2026**	**LR, RF, KNN, DT + PFI + LIME**	**UCI obesity dataset (N = 2,111)**	**Unified comparison with a quantitative local–global consistency analysis**	**Partly synthetic data; single dataset**

3. Methodology

3.1 Dataset Description

For this work, the publicly available Obesity Levels dataset is used; this dataset was taken from the UCI Machine Learning Repository and is available on Kaggle. The Obesity Levels dataset contains 2,111 examples and 17 attributes (16 predictors and one target attribute). Demographic, physiological, and lifestyle factors linked to obesity can be found in the data [19]. The predictors are explained in Table 2.

The target attribute (NObeyesdad) encompasses seven different BMI categories including: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. One distinctive feature of the Obesity Levels dataset lies in its combined nature – 77% of the observations are generated through SMOTE procedure, whereas the rest are real observations [20]. This should be kept in mind when analyzing the results obtained by the models. The data is provided here: https://www.kaggle.com/datasets/fatemehmehrparvar/obesity-levels (another copy is available here: https://www.kaggle.com/datasets/jayitabhattacharyya/estimation-of-obesity-levels-uci-dataset).

Table 2. ***Description of the predictor variables in the Obesity Levels dataset.***

Variable	Description
Gender	Sex of the respondent
Age	Age of the respondent
Height	Height of the respondent
Weight	Body weight of the respondent
family_history_with_overweight	Family history of overweight
FAVC	Frequent consumption of high-caloric food
FCVC	Frequency of vegetable consumption
NCP	Number of main meals per day
CAEC	Consumption of food between meals (snacking habits)
SMOKE	Smoking status
CH2O	Daily water consumption
SCC	Calorie consumption monitoring
FAF	Physical activity frequency
TUE	Time spent using technology devices
CALC	Alcohol consumption

Variable	Description
MTRANS	Mode of transportation usually used

3.2 Implementation Environment

All experiments were conducted in the R programming environment [41], using the caret (Classification and Regression Training) framework for model development, validation, and hyperparameter tuning. The following R packages were used:

- **caret** [42]: unified interface for model training and cross-validation;
- **randomForest** [43]: implementation of the Random Forest classifier;
- **e1071** [44]: supporting statistical and machine learning utilities;
- **rpart** [45]: Decision Tree modeling;
- **pROC** [46]: ROC curve and AUC computation;
- **lime** [47]: local interpretable model-agnostic explanations;
- **ggplot2** [48]: data visualization.

This unified environment ensures consistency across models and reproducibility of the results.

3.3 Data Preprocessing

A comprehensive pre-processing pipeline was used to guarantee data quality, model compatibility, and reproducibility. Missing values from records have been dropped using the na.omit() function of R [41]. The output variable is turned into a categorical factor for multiclass classification. The dataset was split into training (80%) and testing (20%) datasets with stratified sampling by using the createDataPartition() function of the caret package [42]. Stratified sampling ensured that the two classes are proportionately present in both subsets (the generated test dataset consists of 420 records).

For comparability of the features' magnitudes, numerical features are standardized using the preProcess() function of caret [42]. Rescaling was performed for algorithms such as KNN and LR that are sensitive to the scale of the input features. In contrast, tree-based methods were not scaled since they employ splitting criteria based on features that are insensitive to feature scale [49]. A constant value of the random seed (set.seed(123)) has been used across all the experiments [41]. The data resampling has been applied only once during train/test split [49].

3.4 Model Development and Training

Four supervised ML models were developed and trained within the caret framework, which provides a unified interface for model construction, training, validation, and hyperparameter optimization [42]:

- Decision Tree (DT) using the rpart implementation [26], [45];
- Random Forest (RF) using the randomForest package [43];
- Logistic Regression (LR) using multinomial logistic regression [50];
- k-Nearest Neighbors (KNN) classifier [49].

Cross-Validation Strategy

Model training and hyperparameter tuning were performed using five-fold cross-validation through the trainControl() function of caret [42]:

trainControl(method = "cv", number = 5)

Cross-validation is widely adopted to obtain reliable estimates of model performance while reducing the risk of overfitting [49].

Hyperparameter Optimization

Hyperparameter tuning was performed using grid search within the caret framework [42]:

- Decision Tree: the maximum tree depth was varied and further examined through a sensitivity analysis [26];
- Random Forest: the number of trees was fixed at 100 and feature-importance estimation was enabled [43];
- KNN: the optimal number of neighbors (k) was selected through cross-validation [49];
- Logistic Regression: the multinomial model was trained using maximum likelihood estimation [50].

3.5 Model Evaluation Metrics

Evaluation of the predictive performance involved several classification metrics [51]: Accuracy, average F1 score, Cohen's Kappa coefficient [52] and the Area Under the Receiver Operating Characteristic Curve (AUC). The confusion matrix was applied for classification-wise evaluation, while ROC curve was drawn with the One vs Rest approach for each class [46], [51].

3.6 Model Sensitivity Analysis

The sensitivity analysis for the Decision Tree was done using varying depths of the tree [26]. Accuracy was recorded in training and test to find the region of the under-fitting, the optimal complexity of the model, and also overfitting. It is the bias-variance trade-off that is studied in Supervised Learning [49].

3.7 Explainable Artificial Intelligence (XAI) Framework

To improve transparency, a dual-level explainability framework combining global and local explanation methods was adopted [15], [27].

3.7.1 Global Interpretability: Permutation Feature Importance (PFI)

The PFI measure estimates the contribution of each feature to the model performance [27], [53]. PFI estimates the reduction in the predictive performance (in this case, the reduction in the test accuracy) due to random shuffling of the values of one particular feature, thus revealing globally significant features. PFI is model-agnostic, identifies key and secondary features, and allows comparing their importance across different models. However, it can be biased if the features are highly correlated [28].

3.7.2 Local Interpretability: LIME

LIME explains individual predictions by approximating the local behavior of a model with an interpretable surrogate model [15], [47]. It provides instance-level feature contributions, the direction of each feature's influence, and insight into local decision boundaries. In this study, the LIME importance of a feature is the mean absolute LIME weight over the explained instances. These explanations help clarify why a particular prediction was generated, which is especially important in healthcare decision-support systems.

3.7.3 Correlation Analysis Between XAI Methods

For assessing the consistency between the global and local explanations, the PFI and LIME importance scores of the predictors for each model were paired and correlated by means of Pearson's correlation coefficient (r), which evaluates the linear agreement, and Spearman's rank correlation coefficient (ρ), which evaluates the ranking consistency [54].

3.8 Methodological Contribution

The main methodological contribution of this study consists in integrating: (i) several ML models within a single learning scheme; (ii) the dual-level explainability approach with PFI and LIME; and (iii) the correlation-based assessment of explanation consistency. The proposed integrated methodology is expected to enhance both predictive capability and interpretation of ML applications in healthcare.

4. Results

4.1 Class Distribution

Class distribution shows nearly equal values for the seven groups of BMI (Fig. 1), where class ratios go from 12.88% (Insufficient weight, n = 272) to 16.63% (Obesity Type I, n = 351). This fact is favorable for multi-class classification since this way there would be no possibility for the classifier to be biased towards the dominating classes. Nonetheless, the mentioned equality is achieved owing to synthetically created data.

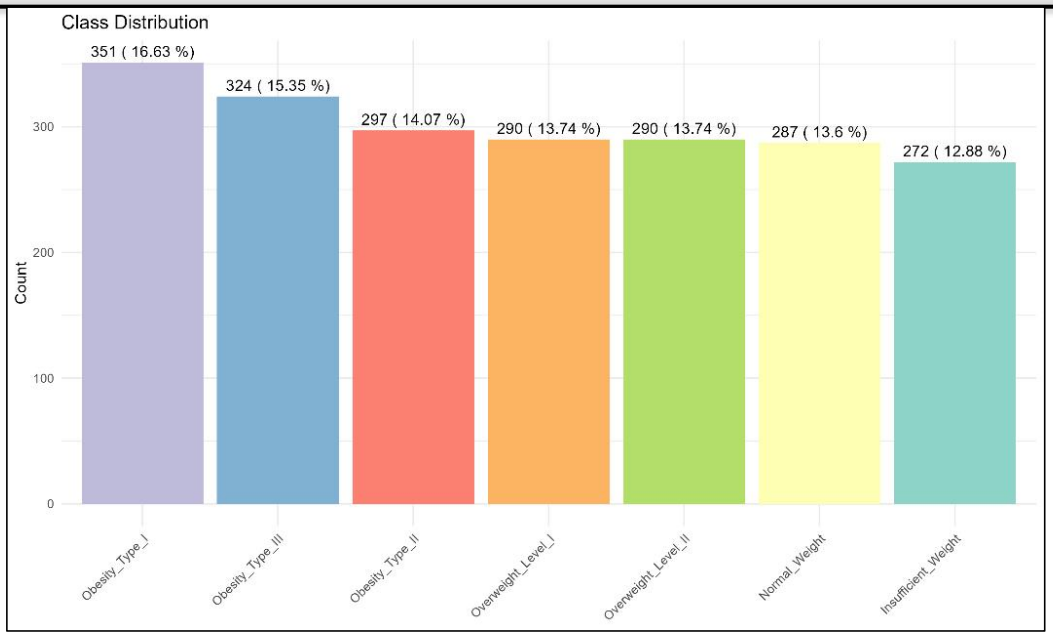


Fig. 1. Distribution of the seven BMI categories in the dataset (N = 2,111).

4.2 Comparative Predictive Performance

The comparison between the four models on the held-out test set is shown in Table 3 and Figures 2 and 3. LR and RF yield the same results in terms of Accuracy (0.962) and Cohen's Kappa (0.956). This shows not only the good results of prediction but also the agreement of the model far beyond the mere level of coincidence. The equality of results indicates that the data is well organized and quite linearly separable.

Table 3. ***Test-set performance of the four models.***

Model	Accuracy	Macro F1	Cohen's Kappa	AUC
Logistic Regression (LR)	**0.962**	**0.961**	**0.956**	0.993
Random Forest (RF)	**0.962**	**0.961**	**0.956**	**0.999**
k-Nearest Neighbors (KNN)	0.812	0.798	0.780	0.940
Decision Tree (DT)	0.545	0.761	0.472	0.850

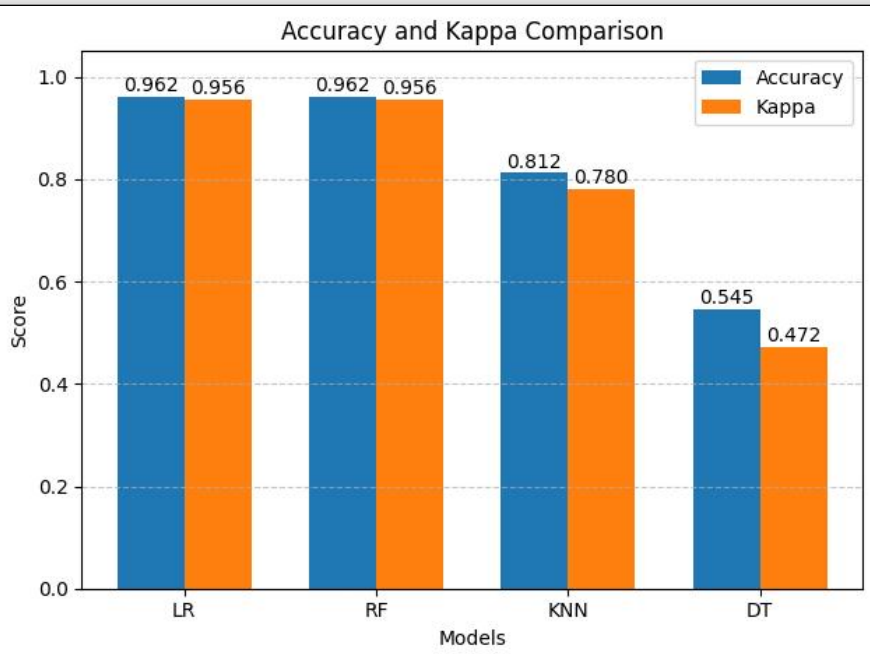


Fig. 2. Comparison of Accuracy and Cohen's Kappa across the four models.

RF slightly outperforms LR in terms of AUC (0.9988 versus 0.993), reflecting a better ability to separate the classes across decision thresholds, and both models maintain high macro F1-scores (0.961), confirming balanced classification across all BMI categories (Fig. 3).

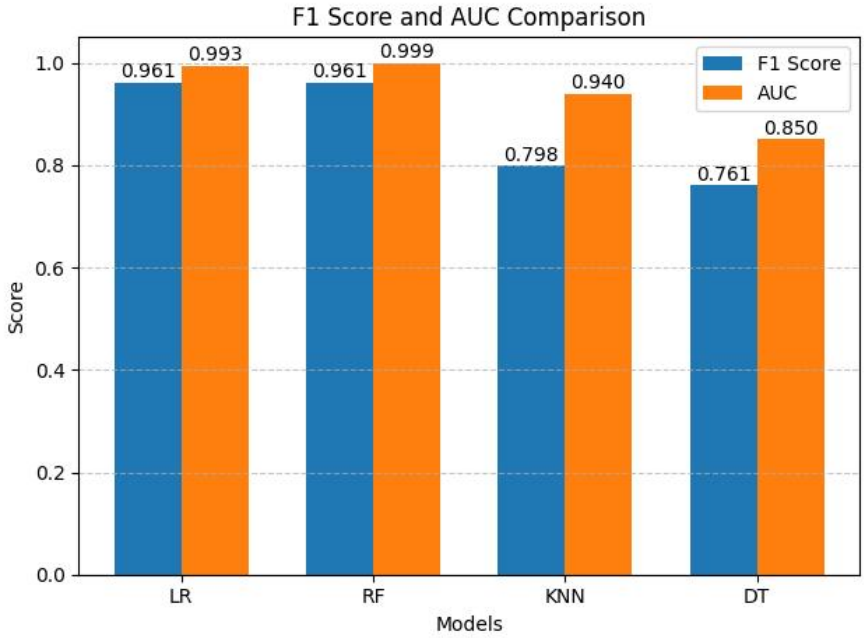


Fig. 3. Comparison of macro F1-score and AUC across the four models.

As seen from both the graphs, the performance of KNN is moderate (Accuracy = 0.812, Kappa = 0.780). This implies that KNN is capable of capturing local features, though its performance suffers when it is asked to generalize in the complex feature space. Clearly there is a performance gap for the decision tree as Accuracy (0.545) and Kappa (0.472) for it are considerably low. Macro F1 score of 0.761 is quite high compared to the accuracy of the decision tree and could happen because the calculation of macro-average is made only over the classes the tree is making predictions about.

4.3 Confusion Matrix of the Random Forest Model

The confusion matrix of the RF model (Fig. 4) shows strong classification performance across all BMI categories, with 404 of the 420 test instances (96.19%) on the diagonal. Obesity Type III is classified perfectly (64/64) and the Obesity Type I and II classes have at most one error each, suggesting that these categories possess distinct feature patterns.

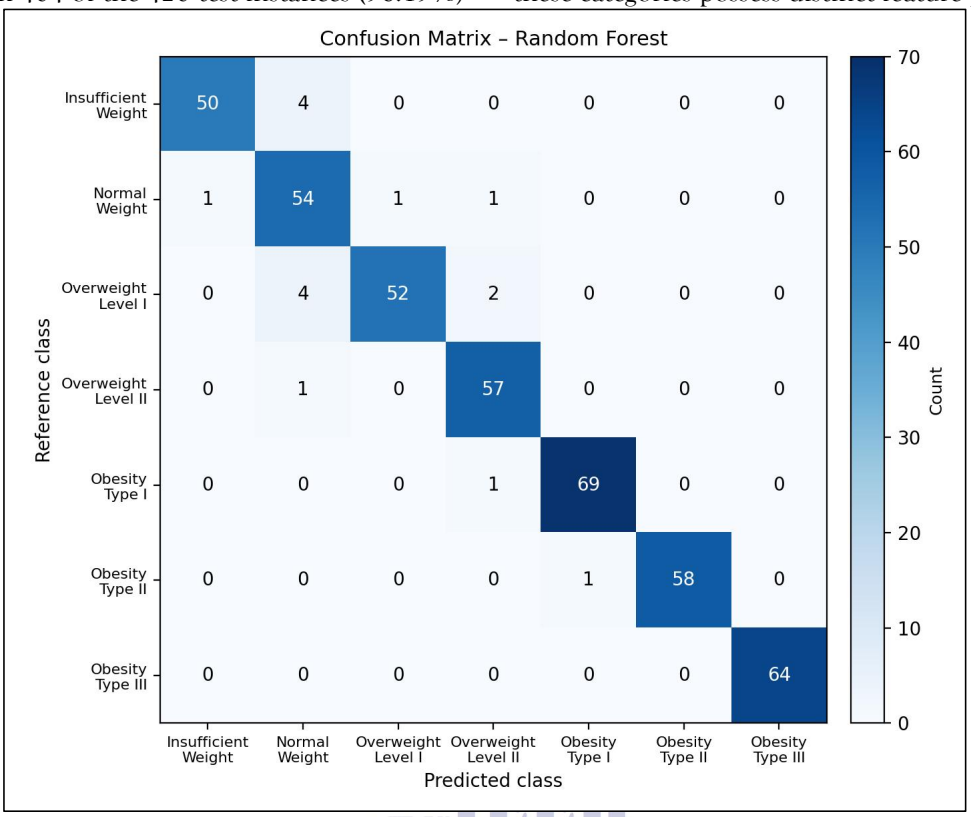


Fig. 4. Confusion matrix of the Random Forest model on the test set (rows: reference class; columns: predicted class).

The confusion matrix of the RF classifier (Fig. 4) demonstrates very good classification accuracy among all BMI groups, with 404 out of 420 instances in the test set (96.19%) lying on the diagonal. The Obesity Type III category is recognized without any errors (64/64), and the Obesity Type I and II categories contain at most one error each. Thus, it can be concluded that these categories have distinctive feature sets.

All 16 incorrect classifications are clustered around the class borders, and 15 of them belong to adjacent BMI categories: for instance, Insufficient Weight classified as Normal Weight (4 cases), Overweight Level I classified as Normal Weight (4 cases) or Overweight Level II (2 cases), and Normal Weight classified as Overweight Level I (1 case). This systematic behavior instead of random confusion suggests that the ordinality of BMI categories is correctly captured by the classifier, but borderline cases with overlapping feature distributions still constitute the source of mistakes.

4.4 Decision Tree Depth Sensitivity

The graph in Fig. 5 illustrates the connection between complexity and accuracy of the DT model. The model experiences underfitting at low depths (1-3) since the accuracy increases sharply from 0.29 at depth 1 to 0.65 at depth 3 due to the inability of the tree to capture the patterns. The accuracy further increases up to depth 6 after which the two graphs plateau at 0.88 (training) and 0.86 (testing) respectively. The small and constant difference between the two graphs implies no overfitting.

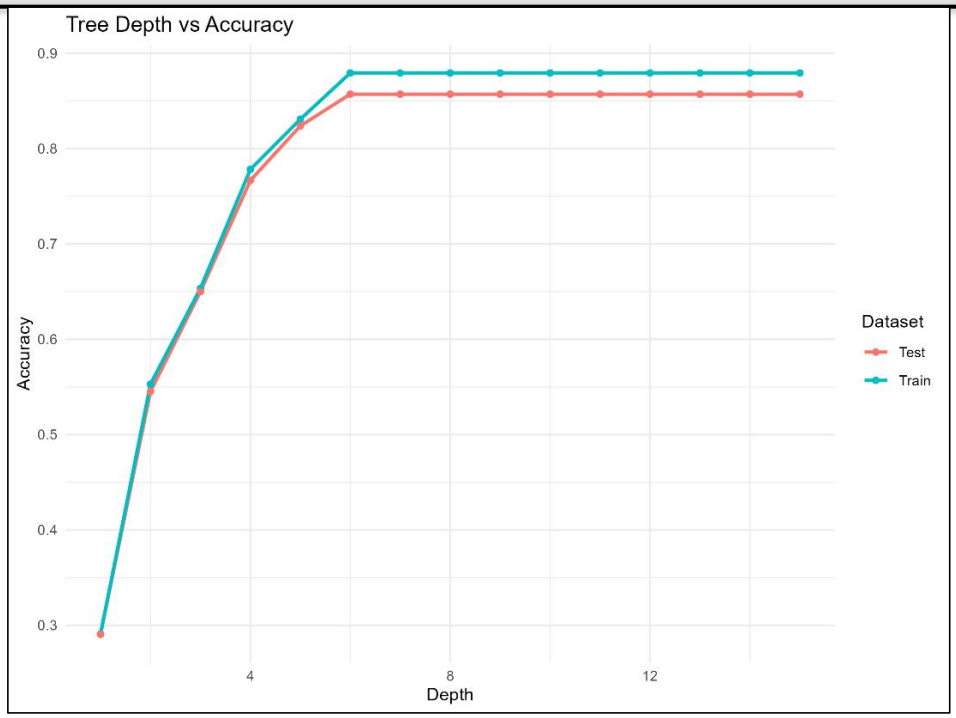


Fig. 5. Effect of tree depth on the training and testing accuracy of the Decision Tree.

4.5 Global Interpretability: Permutation Feature Importance

To identify the key predictors of obesity class, PFI was computed for DT, KNN, LR, and RF as the drop in test accuracy after permuting each feature (Figs. 6–9; each figure shows the ten highest-ranked features).

Decision Tree (DT)

From Figure 6, it is evident that the DT has a very high concentration of importance. Weight is the most important variable because its permutation causes a decrease in accuracy of 0.381, followed by Gender with 0.124 importance. All other variables have zero importance. It is worth noting that the Height variable, together with the Weight variable, is responsible for determining the BMI, but it gets zero importance.

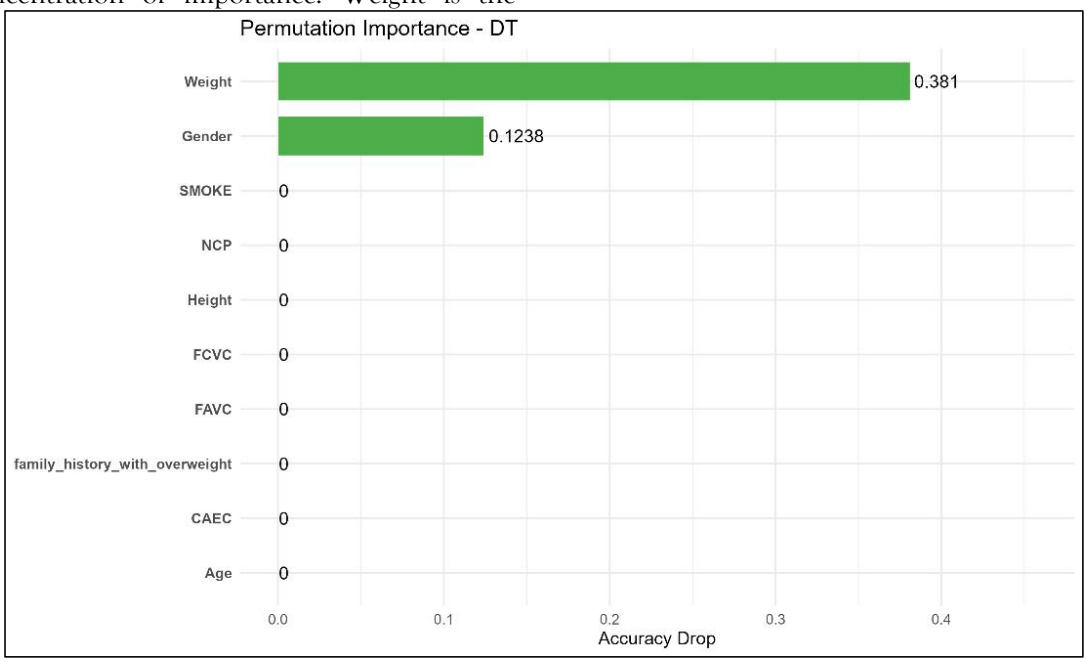


Fig. 6. Permutation feature importance (accuracy drop) for the Decision Tree: top 10 features.

k-Nearest Neighbors (KNN)

On the other hand, the KNN model possesses a distributed importance distribution profile (Figure 7). Weight continues to be the most important variable (0.143), however, there are other variables which make significant contributions as well: Gender (0.114), CALC (0.114), TUE (0.076), CAEC (0.069), and Age (0.069). Family history, FCVC, MTRANS, and FAVC are the other moderately contributing variables (0.055 – 0.062). The pattern is due to the inherent characteristic of KNN which makes it an instance-based learner and relies on distance within multi-dimensional feature space.

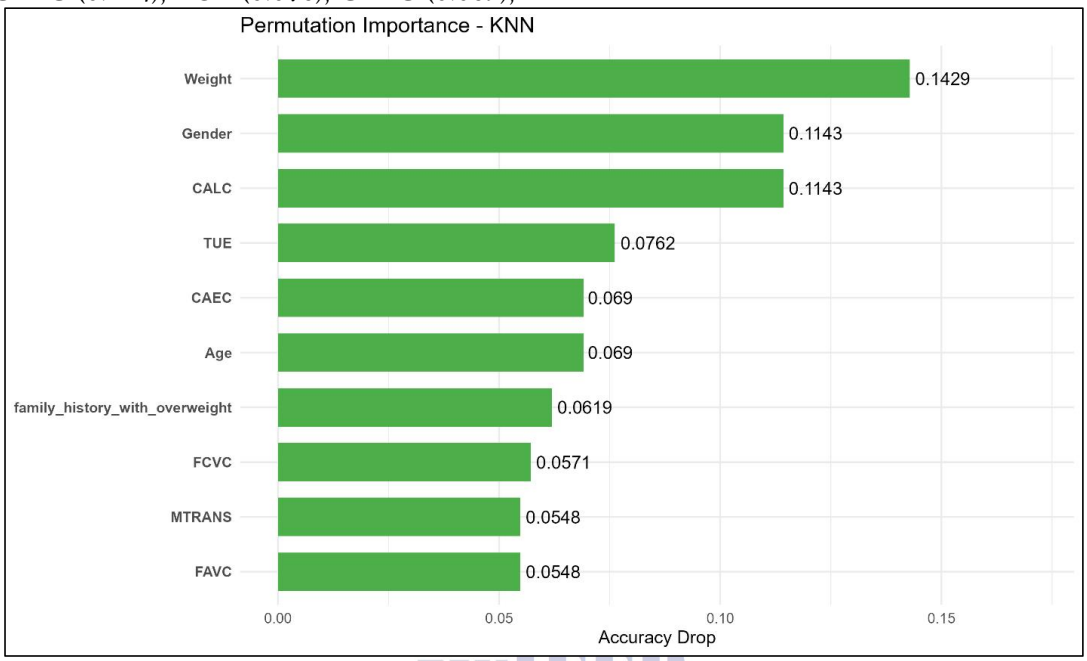


Fig. 7. Permutation feature importance (accuracy drop) for k-Nearest Neighbors: top 10 features.

Logistic Regression (LR)

The LR importance distribution (Figure 8) is the one that has the most physiological interpretation. Weight is the parameter with the greatest importance (0.769), followed by Height (0.352), which corresponds to the definition of BMI. The other factors contributing to the importance distribution are: Gender (0.136), frequency of performing physical activities FAF (0.045), frequency of consuming vegetables FCVC (0.043) and Age (0.041), as well as lesser contributions from CALC (0.026), CAEC (0.024), SCC (0.014) and CH2O (0.010).

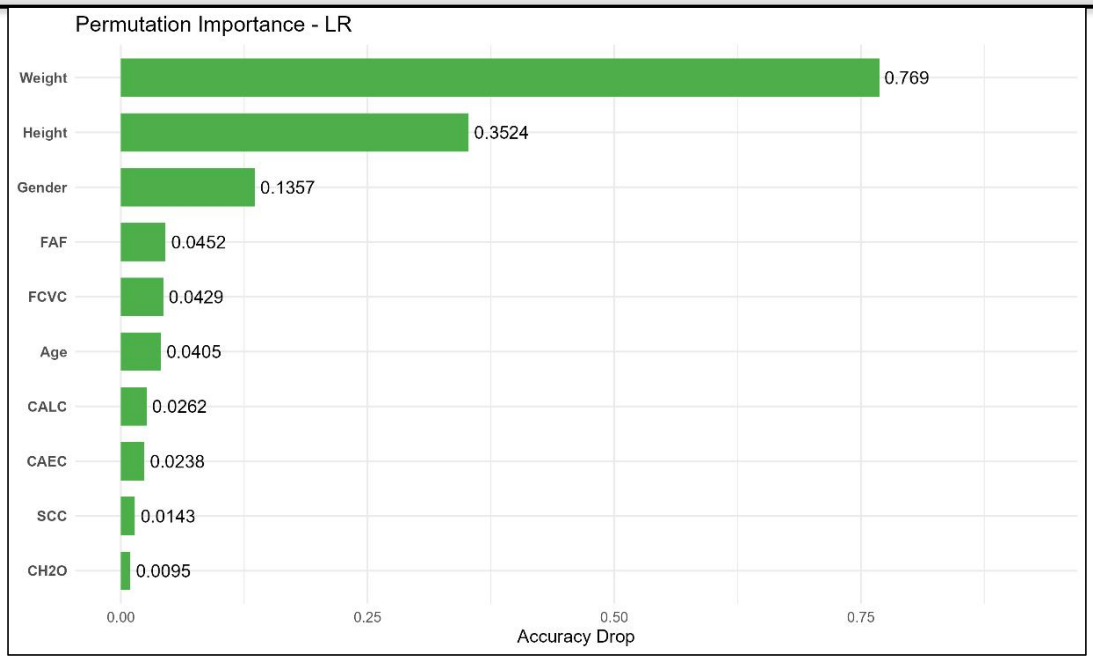


Fig. 8. Permutation feature importance (accuracy drop) for Logistic Regression: top 10 features.

Random Forest (RF)

The RF model (see Fig. 9) demonstrates a hierarchically structured importance of features. The most important predictors are Weight with an accuracy reduction of 0.748 and Height (0.212). Moderately important predictors are FCVC (0.117), Gender (0.086), and Age (0.036). Less important but still valuable features include FAVC (0.017), CAEC (0.010), having family history with overweight (0.007), and CH2O (0.007). Features like TUE, NCP, SCC, FAF, and CALC have little importance, and SMOKE and MTRANS do not affect the accuracy at all. Such hierarchy implies that RF takes into account both major physiological factors and additional secondary lifestyle factors due to its ensemble nature, which involves multiple trees.

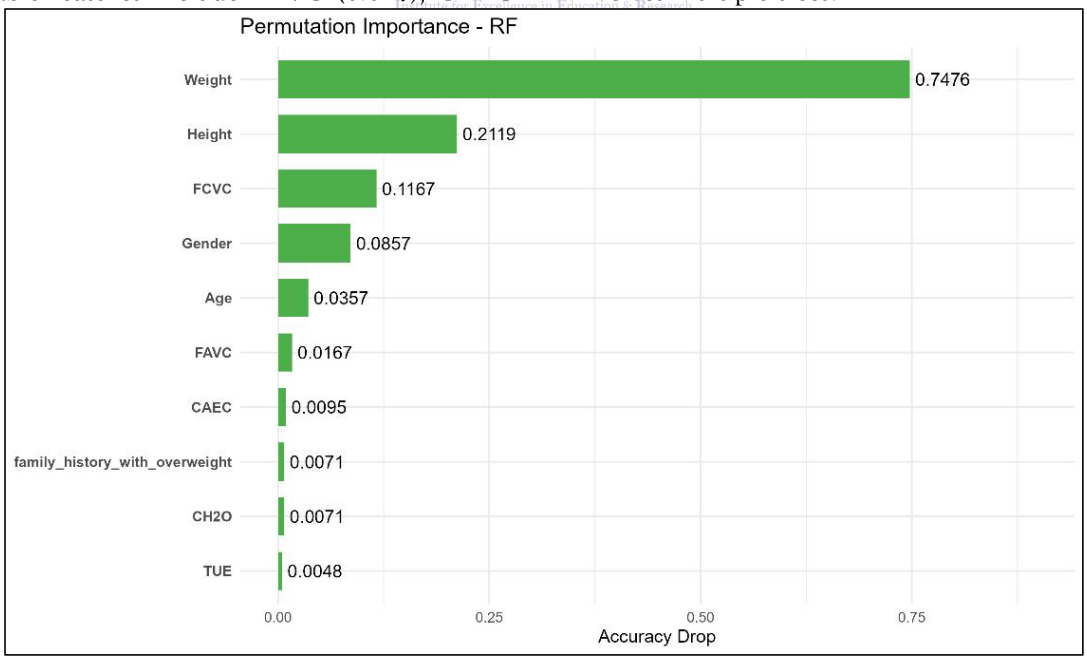


Fig. 9. Permutation feature importance (accuracy drop) for Random Forest: top 10 features.

4.6 Local Interpretability: LIME

In addition to the global insights obtained from PFI, LIME was applied to gain local explanations on how each of the four models made their predictions (Figs. 10-13; mean absolute LIME weight, top 10 features). Unlike PFI, which uses performance reduction to

determine the global relevance of variables, LIME describes the impact that feature contribution has on individual prediction.

Decision Tree (DT)

The LIME values obtained for the DT suggest that the attribution process is very concentrated, similar to the results obtained using PFI (Fig. 10). The main variable in the model is weight with the importance score of 0.276, followed by Gender (0.091), while all other variables have almost no importance (0.012 or less). It appears that the DT is dependent on decision-making boundaries that are localized and determined by weight splits only.

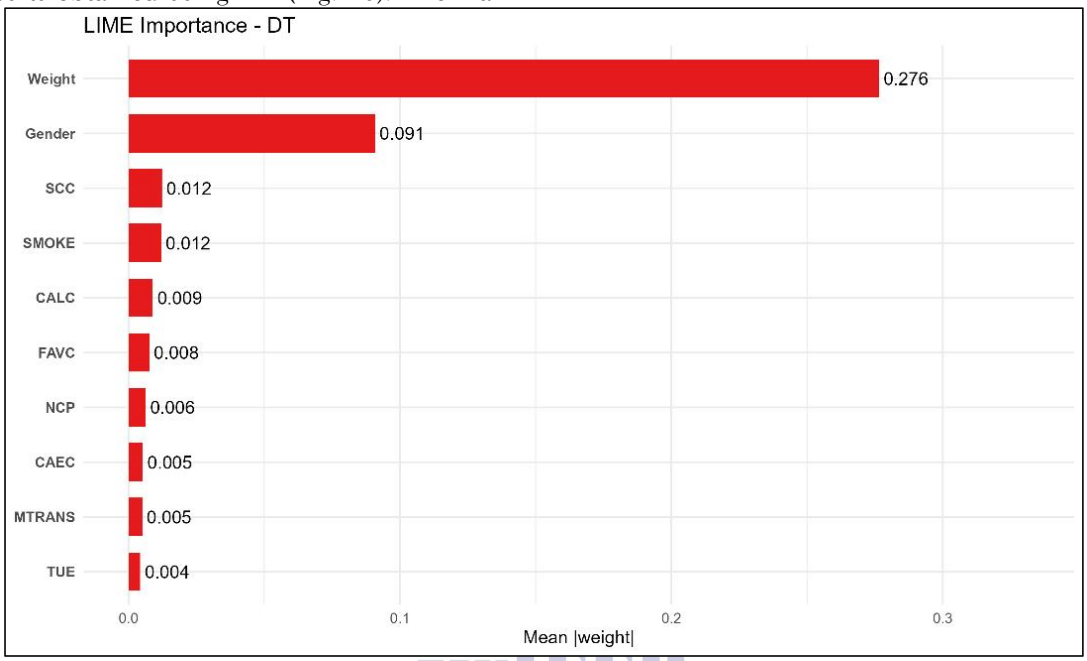


Fig. 10. LIME feature importance (mean absolute weight) for the Decision Tree: top 10 features.

k-Nearest Neighbors (KNN)

However, the KNN algorithm presents a highly distributed and behaviorally driven importance distribution in its local explanations based on LIME (see Fig. 11). The main features that affect the outcome the most are SMOKE (0.172), CAEC (0.136), SCC (0.129), family history of being overweight (0.112), and FAVC (0.107), while the second most important feature is Gender (0.082). Unlike in the PFI approach, Weight (0.068) does not play the key role here. The new situation demonstrates that KNN is largely affected by the local neighborhood structure, which includes the behavioral similarity more than physical characteristics.

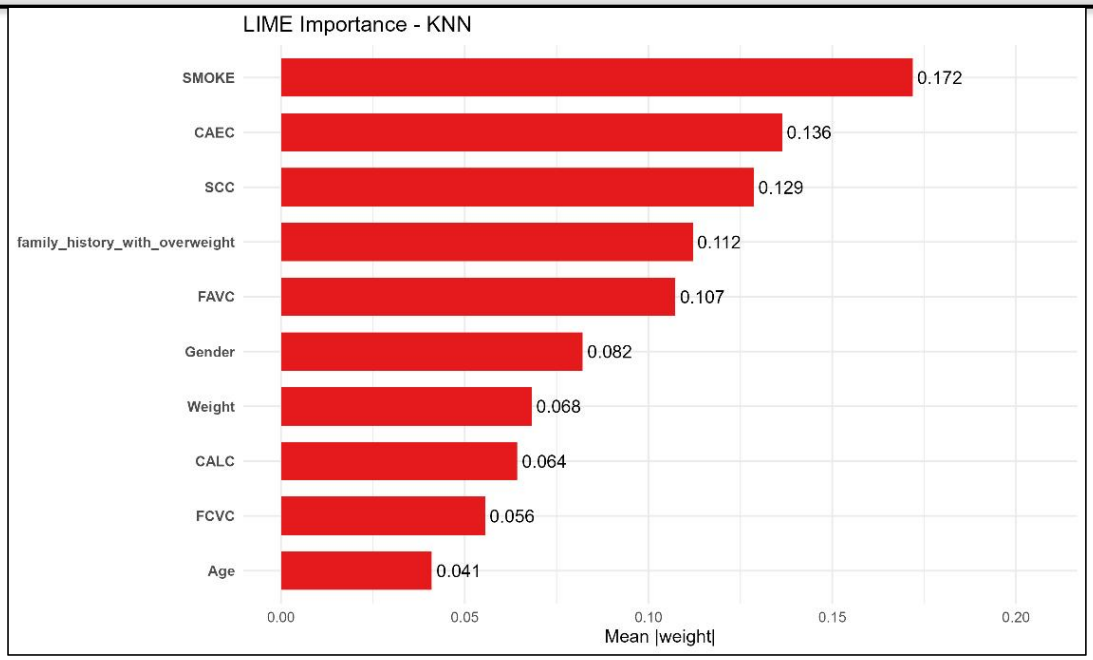


Fig. 11. LIME feature importance (mean absolute weight) for k-Nearest Neighbors: top 10 features.

Logistic Regression (LR)

The LIME analysis for LR reveals a balanced and clinically meaningful attribution profile (Fig. 12). The Weight is the most prominent feature (0.265), followed by Age (0.088), SCC (0.085), and Gender (0.076). Secondary contributions can be identified for Height (0.033), CALC (0.032), SMOKE (0.032), and CAEC (0.030). Contrary to PFI, where the LR model focused on the BMI-related features of Weight and Height, LIME highlights the fact that the model's predictions involve both physiological and behavioral factors. This way, LR preserves global linear consistency and still allows local variability of feature importance.

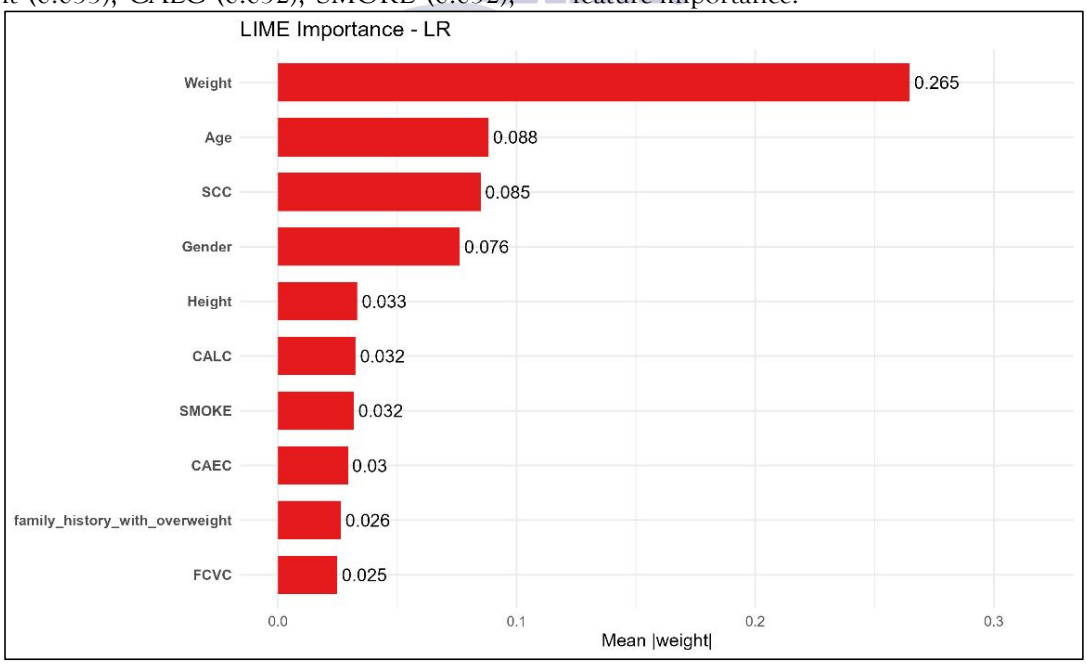


Fig. 12. LIME feature importance (mean absolute weight) for Logistic Regression: top 10 features.

Random Forest (RF)

The hierarchy of LIME explanation for RF is still observable (Fig. 13). Weight remains the most important factor (0.264), while Gender (0.055), family history with overweighting (0.037), Height (0.034), FCVC (0.030), and Age (0.027) are moderately important. Behavioral factors like FAVC (0.024), CAEC (0.018), and SCC (0.016)

demonstrate medium importance. The LIME explanation generated by RF is much more diversified than that generated by DT, as it better reflects the complex behavior of the random forest; on the other hand, it is still hierarchically structured similarly to KNN's output.

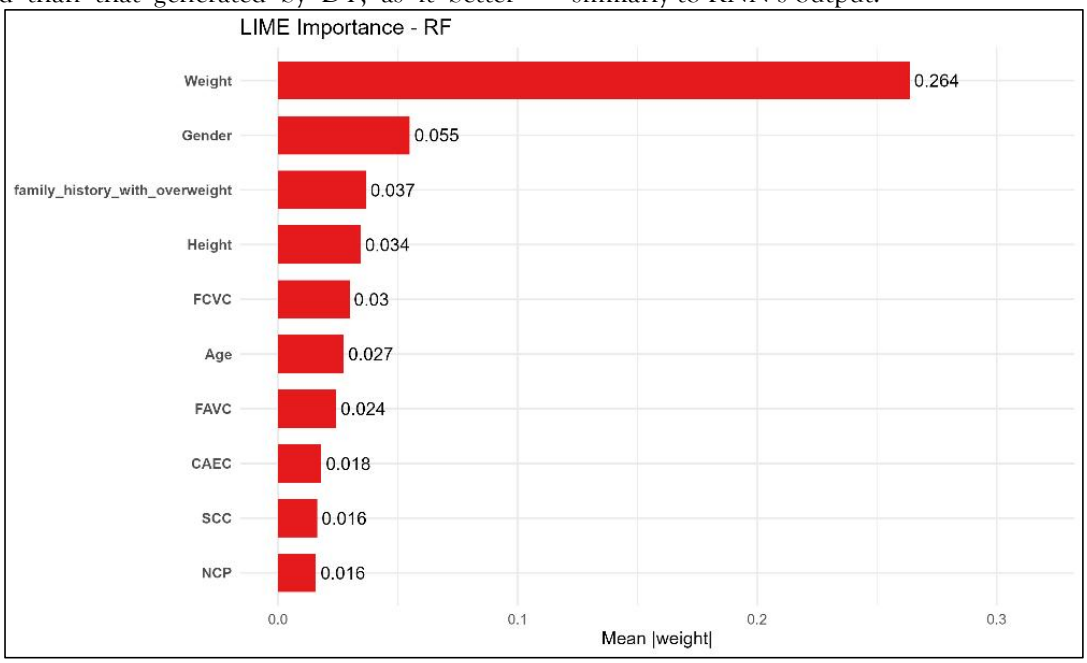


Fig. 13. LIME feature importance (mean absolute weight) for Random Forest: top 10 features.

4.7 Consistency Between LIME and Permutation Importance

To investigate the consistency between the global (PFI) and local (LIME) explanations, Pearson and Spearman correlations were computed between the importance scores of the two methods for each model (Table 4; Figs. 14–17).

Table 4. ***Agreement between LIME and PFI importance scores.***

Model	Pearson *r*	Spearman *ρ*	Interpretation
DT	0.999	0.592	Near-perfect linear agreement, inflated by Weight (and Gender); many tied zero-importance features
KNN	−0.106	0.097	No meaningful agreement; strong local–global divergence
LR	0.854	0.564	Strong linear but moderate rank agreement; driven largely by Weight, with Height diverging
RF	0.969	0.817	Strong linear and highest rank agreement among the models

Decision Tree (DT)

As demonstrated in Fig. 14, there is a nearly perfect linear relationship between LIME and PFI in case of the DT (r = 0.999), while Spearman's ρ value is only moderate (ρ = 0.592). It is due to the scale-induced coincidence and not the actual coincidence in terms of rankings, because Weight and Gender are the dominant factors in both cases, while there is almost a perfect linear relationship between them and all other factors are situated close to zero in both cases. Therefore, Pearson's value is exaggerated because of the predominance of a single factor, while the Spearman value is limited by numerous zeros.

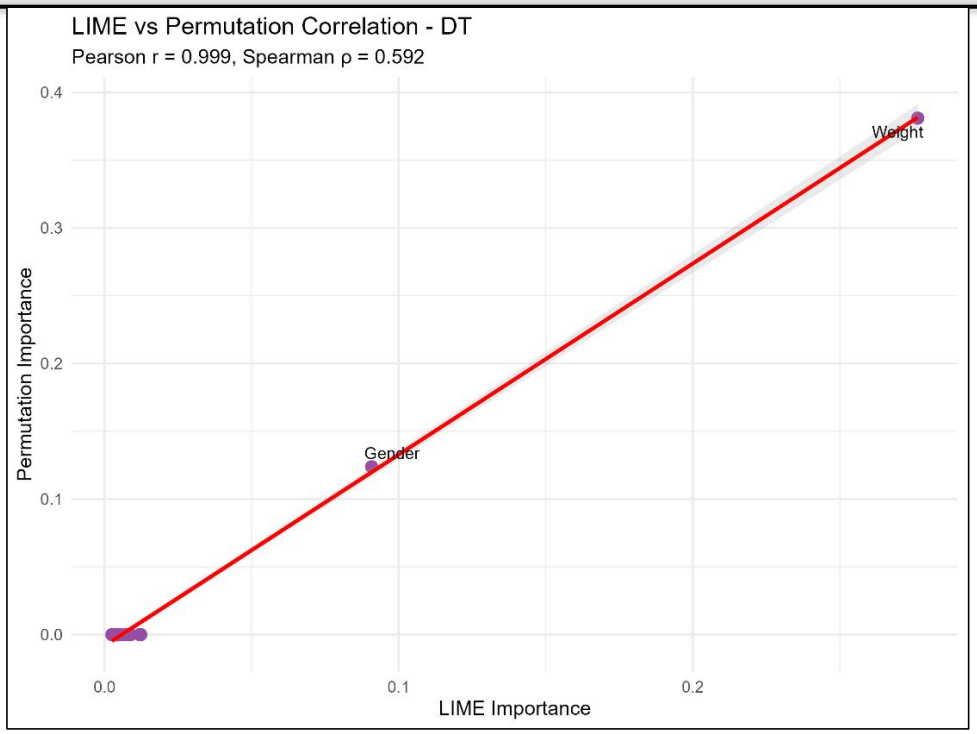


Fig. 14. Correlation between LIME and permutation importance for the Decision Tree (Pearson r = 0.999, Spearman ρ = 0.592).

k-Nearest Neighbors (KNN)

As can be seen from Fig. 15, for KNN there is no correlation between LIME and PFI (r = −0.106, ρ = 0.097). The scatter is very dispersed and nonlinear, which means that there is a big difference between global and local importance, since SMOKE and SCC have high LIME importance but zero PFI importance, meaning that these features have a great local importance but a low global one, while Weight, CALC, and Gender have greater importance globally than locally. It is due to the instance-based nature of KNN, where LIME gives explanations about behavior specific to instances, while PFI provides average global importance of features.

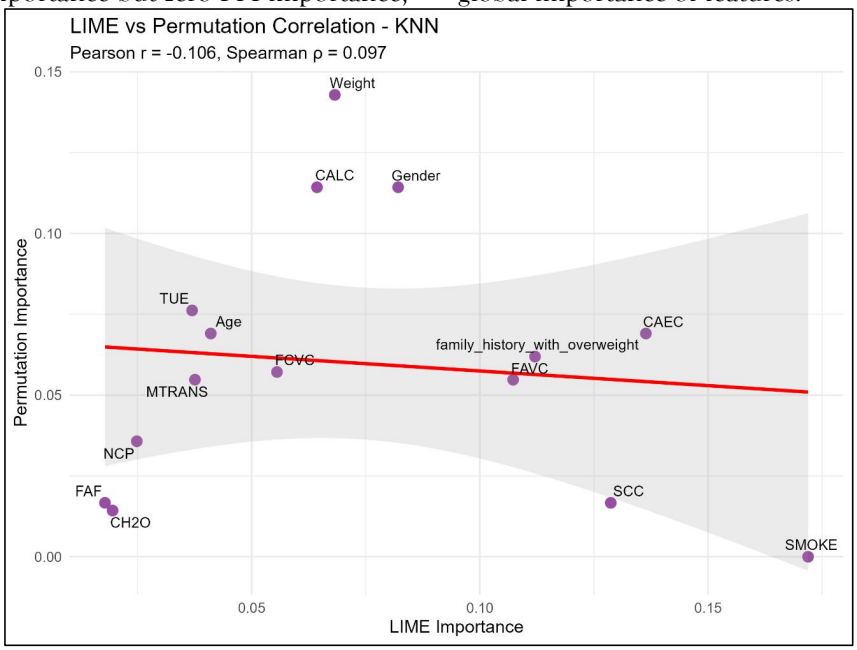


Fig. 15. Correlation between LIME and permutation importance for k-Nearest Neighbors (Pearson r = −0.106, Spearman ρ = 0.097).

Logistic Regression (LR)

From Fig. 16, we see that the LR model exhibits a very high positive linear correlation between LIME and PFI (r = 0.854) and a medium rank correlation (ρ = 0.564). Weight is ranked first in both methods, and Gender has a constant moderate importance, while Height is an outlier since it is the second most important feature in PFI (0.352) and gets a small weight in LIME (0.033). Also, Age and SCC get higher ranking in LIME compared to PFI. Features with low ranking (FAVC, SMOKE, NCP) are clustered around zero in both methods. The Pearson correlation is very sensitive to the high-leverage point of Weight, while the moderate Spearman correlation is the more conservative metric since LR maintains the relative ranking of its key features but does not preserve the exact rankings.

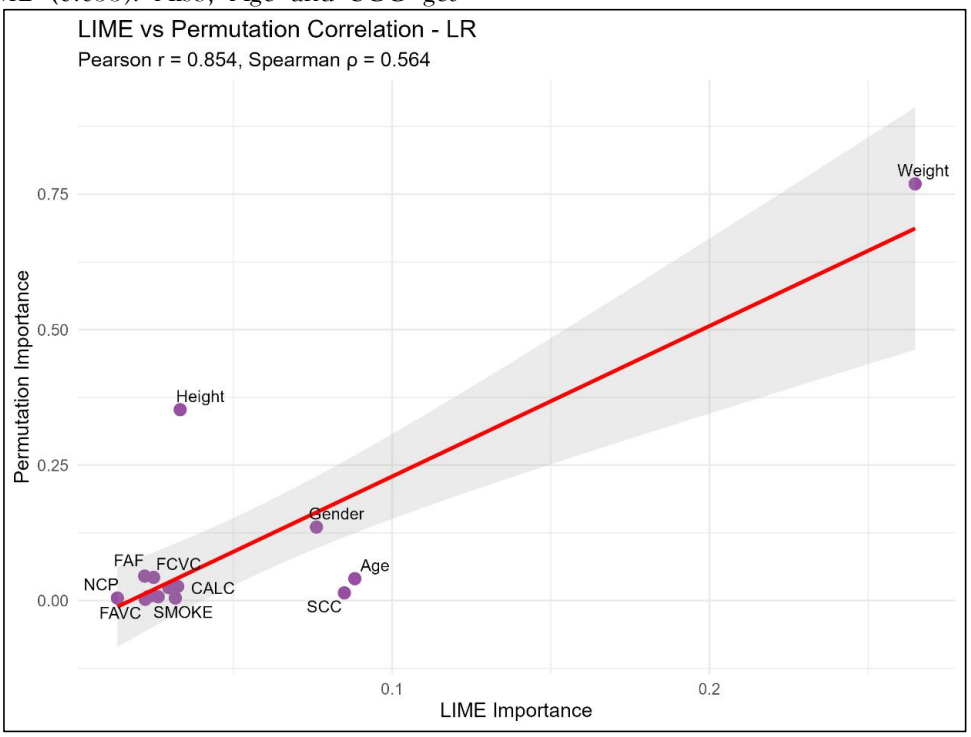


Fig. 16. Correlation between LIME and permutation importance for Logistic Regression (Pearson r = 0.854, Spearman ρ = 0.564).

Random Forest (RF)

According to Fig. 17, for the RF model, there is very high linear dependence between LIME and PFI (r = 0.969) and highest agreement in ranking (ρ = 0.817). Weight is a dominant predictor with aligned rankings; Gender and Age have moderate impact with some discrepancy in ranking. Height and FCVC are above the regression line, meaning that their impact is higher than predicted locally and globally (LIME vs PFI). Lower-ranking predictors like TUE are clustered around the lower part of both scales. The high value of Pearson's r shows that the levels of impact have been preserved, while the value of Spearman's ρ shows that the rankings are very consistent, even though not perfectly so. In contrast to the DT model, for which the correlation was biased by one predictor, the RF model is more consistent in its feature importances.

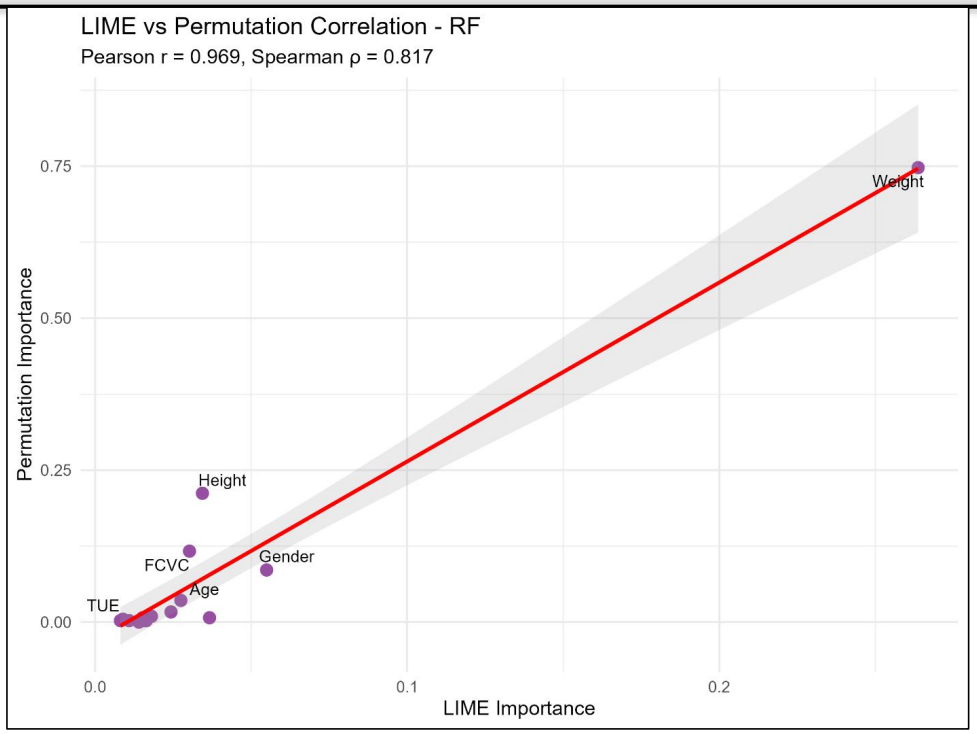


Fig. 17. Correlation between LIME and permutation importance for Random Forest (Pearson r = 0.969, Spearman ρ = 0.817).

4.8 Cross-Model Synthesis of Global and Local Explanations

Table 5 summarizes the three highest-ranked features of each model under PFI and LIME. Integrating the two methods yields the following picture.

Table 5. *Top-three features per model under PFI (accuracy drop) and LIME (mean absolute weight).*

Model	PFI (global)	LIME (local)
DT	Weight (0.381), Gender (0.124); all other features 0	Weight (0.276), Gender (0.091), SCC (0.012)
KNN	Weight (0.143), Gender (0.114), CALC (0.114)	SMOKE (0.172), CAEC (0.136), SCC (0.129)
LR	Weight (0.769), Height (0.352), Gender (0.136)	Weight (0.265), Age (0.088), SCC (0.085)
RF	Weight (0.748), Height (0.212), FCVC (0.117)	Weight (0.264), Gender (0.055), Family history (0.037)

Consistent core predictor

Weight is the key factor in PFI for all four models and in LIME for DT, LR, and RF. The only exception is KNN, where under LIME the first three local factors are SMOKE, CAEC, and SCC, and Weight ranks only seventh. Thus, Weight becomes the key physiological factor in most models in predicting obesity, while Height is the key factor in LR and RF (and is not considered at all in DT).

Divergence in behavioral features

Behavioral variables show moderate global importance under PFI but can become locally dominant under LIME, particularly for KNN (SMOKE, CAEC, SCC, FAVC), which indicates strong instance-specific influence and sensitivity to individual lifestyle patterns.

Model-specific interpretability patterns

- **DT:** overly simplistic in both global and local contexts; dominated by one or two features, with low explanatory depth.
- **KNN:** weak global prioritization (PFI) but strong local variability (LIME); captures contextual and behavioral heterogeneity at the cost of consistency.
- **LR:** strong global physiological grounding (PFI) and balanced local explanations (LIME); the best alignment with clinical interpretability.
- **RF:** balanced global importance (PFI) and structured local explanations (LIME); the best trade-off between stability and complexity.

Key insight

Classification of obesity in this dataset does not depend on the weight alone but involves a multi-faceted approach, where Weight and Height form the basis of physiological (BMI-based) classification, while diet and behavioral variables form the classification within that basis. The Weight variable serves as the primary anchor for classification while behavioral/lifestyle variables act as modifiers locally. This underscores the context-dependency of obesity classifications in which people with the same weight characteristics are placed in different classes depending on dietary behavior, physical activity, etc. The results are in line with accepted clinical knowledge and thus lend credibility to the external validity of the models, even though the data is synthetic.

5. Discussion

Among others, there are some conclusions made about model performances and their explainability in classifying obesity. First, LR and RF provided the best performance in terms of all metrics used (Accuracy = 0.962, Kappa = 0.956), hence demonstrating a high effectiveness of both linear and ensemble models for this data set. Equivalence of these models indicates some level of linear separability of the data, although the RF model provides better discriminative capabilities (AUC = 0.9988). There were higher accuracies obtained for the same data set using stacked ensembles [29]. Thus, the novelty of this study does not lie in the obtained accuracy rate but in its explainability.

The second observation from the tree depth analysis is that the DT achieves a test accuracy of around 0.86 when the tree depth is allowed to go up to 6, while the DT in the model comparison achieved only 0.545. The last value corresponds to the accuracy of the very shallow tree in Fig. 5 (with a depth of 2), and is consistent with the PFI result which showed that the tuned tree utilizes only Weight and Gender. Thus, the poor DT performance in Table 3 should be attributed to the complexity of the tuned tree rather than the limitation of tree learners.

Third, from the confusion matrix, the errors made by RF are not random but are concentrated between adjacent BMI classes. In other words, the classifier understands the ordered structure of the obesity classes, yet it performs poorly when feature distributions are mixed.

On the other hand, interpretability results provide substantial support that classification is based on a multi-dimensional feature structure. Weight is the most important feature under PFI for all classifiers and under LIME for DT, LR, and RF due to its significance in BMI-based classification. The Height feature is only significant for LR and RF. Some behavioral and lifestyle features like diet, physical activities, and eating habits are essential in local classification, especially in case of KNN where LIME explanation relies on SMOKE, CAEC, and SCC.

The comparison of LIME and PFI also demonstrates the difference between the global and local interpretability of ML models. While PFI provides information about the overall feature importance, LIME allows to see the differences within instances and to show that each prediction is made depending on some context. The correlation analysis provides another perspective on the difference between these two approaches – there is high correlation of the two measures in RF (ρ = 0.817), moderate in LR (ρ = 0.564) and none in KNN (ρ = 0.097). The perfect Pearson correlation coefficient of the DT model is due to one important feature dominating all the others. As Weight is much farther from other

features in the scatter plots of LR and RF, the Pearson correlation coefficients for these models should also be considered along with Spearman coefficients. In general, RF is the model that provides the best trade-off between predictivity and interpretability, allowing to get stable global and local explanations, while LR provides the most clinically interpretable structure.

In general, the study shows that the combination of predictive models and complementary explanations leads to the improvement of the reliability and interpretability of ML systems in healthcare. Relatively simple models can reach high performance when their explanations are examined at both the global and the local level.

6. Limitations

However, it is also necessary to outline certain limitations despite such positive results. First of all, a rather large share of the dataset includes synthetic examples, that can bring bias and result in inflated model performance [19], [20], which makes it difficult to apply this work to real people. Secondly, only one single dataset was used in the research, which means that there are no means of testing the model's robustness for different demographics. Lastly, since the predictors "Weight" and "Height" are dominant, this could cause feature redundancy as these are the defining parameters of the BMI, and, therefore, make other predictors less significant. Fourth, while it is true that LIME generates useful local explanations, its local approximation tends to generate inconsistencies among instances that are alike, and PFI tends to be biased when there are correlated features [28]. There was no investigation into global explanation techniques other than PFI. Fifth, the analysis is based only on one train-test data split without multiple resampling techniques or confidence intervals and tests for significance. Also, the LIME-PFI correlations are estimated using few variables and are susceptible to outliers like weight. Lastly, no advanced model like deep learning was applied here.

7. Conclusion

In this study, a detailed analysis of machine learning models for multi-class obesity prediction was done considering performance measures along with explainability methods. LR and RF performed best (Accuracy = 96.19%, Kappa = 0.956) with RF giving the highest AUC (0.9988). The Decision Tree is inherently explainable; however, it did not perform well since the tuned tree considered only Weight and Gender. On the other hand, KNN performed well by considering behavior patterns but lacked consistency.

The feature importance analysis indicates that Weight is the most prominent predictor in all models according to PFI, Height also plays an important role for LR and RF, while behavioral features are used for additional precision and become locally important for KNN. The combination of LIME and PFI along with the correlation analysis allows creating a multi-level framework for model interpretation: RF demonstrates the highest level of consistency between local and global explanations, while LR offers the most clinically meaningful structure. In general, the research emphasizes the need to combine accuracy with interpretability when developing ML models in healthcare, while pointing out the artificial nature of the dataset.

8. Future Work

The above study opens up a range of opportunities for future research in various ways. First, the evaluation of the models on several real-life datasets will increase the robustness and generalize the results across different populations. Second, using other explainable algorithms such as SHAP [56] to gain the comprehensive picture regarding how the model works by looking at both global and local explanations would be possible, which will allow the investigation of LIME-PFI consistency to be done on other explanation algorithms as well. Third, using more diverse and enriched data such as clinical notes, wearable sensor data, and genetic data could help to better predict and assess the risk of obesity. Fourth, developing hybrid and deep learning models while keeping the interpretability is an option that might be taken into consideration. Fifth, the use of

repeated cross-validation and statistical tests would make it possible to test the difference between the models and between explanation approaches within confidence intervals. Finally, deployment of the suggested framework in real-world decision support systems is needed.

References

[1] Z. Abad-Jiménez and T. Vezza, "Obesity: A global health challenge demanding urgent action," *Biomedicines*, vol. 13, no. 2, p. 502, 2025.

[2] H. J. Lim, H. Xue, and Y. Wang, "Global trends in obesity," in *Handbook of Eating and Drinking: Interdisciplinary Perspectives*. Springer, 2020, pp. 1217–1235.

[3] K. D. Hesketh, M. Zheng, and K. J. Campbell, "Early life factors that affect obesity and the need for complex solutions," *Nat. Rev. Endocrinol.*, vol. 21, no. 1, pp. 31–44, 2025.

[4] M. H. Wang, "Artificial intelligence across the obesity continuum: From mechanistic insights to global precision prevention and therapy," *Obesity*, vol. 34, no. 2, pp. 294–316, 2026.

[5] J. Rahnenführer *et al.*, "Statistical analysis of high-dimensional biomedical data: A gentle introduction to analytical goals, common approaches and challenges," *BMC Med.*, vol. 21, no. 1, p. 182, 2023.

[6] S. Amjad *et al.*, "Genomic insights: Predicting obesity through AI and machine learning," *Complement. Altern. Med.: Non-Conventional*, vol. 90, 2024.

[7] A. Chatterjee, M. W. Gerdes, and S. G. Martinez, "Identification of risk factors associated with obesity and overweight—A machine learning overview," *Sensors*, vol. 20, no. 9, p. 2734, 2020.

[8] L. Huang *et al.*, "The role of artificial intelligence in obesity risk prediction and management: Approaches, insights, and recommendations," *Medicina*, vol. 61, no. 2, p. 358, 2025.

[9] T. A. A. Abdullah, M. S. M. Zahid, and W. Ali, "A review of interpretable ML in healthcare: Taxonomy, applications, challenges, and future directions," *Symmetry*, vol. 13, no. 12, p. 2439, 2021.

[10] M. Ennab and H. Mcheick, "Enhancing interpretability and accuracy of AI models in healthcare: A comprehensive review on challenges and future directions," *Front. Robot. AI*, vol. 11, p. 1444763, 2024.

[11] M. Hassan, A. Kushniruk, and E. Borycki, "Barriers to and facilitators of artificial intelligence adoption in health care: Scoping review," *JMIR Hum. Factors*, vol. 11, p. e48633, 2024.

[12] W. Shafik, A. F. Hidayatullah, K. Kalinaki, H. Gul, R. Y. Zakari, and A. Tufail, "A systematic literature review on transparency and interpretability of AI models in healthcare: Taxonomies, tools, techniques, datasets, open research challenges, and future trends," *Health Technol.*, pp. 1–22, 2026.

[13] A. Khan, A. Ali, M. I. Mohmand, M. Zareei, and R. R. Biswal, "Explainable artificial intelligence in software engineering: Current trends, gaps, and future directions," *IEEE Access*, 2026.

[14] T. Hulsen, "Explainable artificial intelligence (XAI): Concepts and challenges in healthcare," *AI*, vol. 4, no. 3, pp. 652–666, 2023.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[16] J. Allgaier, L. Mulansky, R. L. Draelos, and R. Pryss, "How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare," *Artif. Intell. Med.*, vol. 143, p. 102616, 2023.

[17] N. S. Alsulami, M. S. Ramzan, B. A. Alzahrani, and S. S. Alsulami, "Obesity prediction using an explainable deep learning framework based on LSTM–LIME with integrated visualization," *Sci. Rep.*, 2025.

[18] G. Singh, N. Chaudhary, P. Verma, R. Kler, A. Gupta, and D. Ather, "Exploring and modeling factors influencing obesity with a comprehensive

analysis using synthetic and real-world data," in *Proc. 2024 Int. Conf. Intell. Innov. Pract. Eng. Manag. (IIPEM)*, IEEE, 2024, pp. 1–9.

[19] F. M. Palechor and A. de la Hoz Manotas, "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico," *Data Brief*, vol. 25, p. 104344, 2019, doi: 10.1016/j.dib.2019.104344.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[21] D. R. Cox, "The regression analysis of binary sequences," *J. R. Stat. Soc. B (Methodol.)*, vol. 20, no. 2, pp. 215–232, 1958, doi: 10.1111/j.2517-6161.1958.tb00292.x.

[22] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[23] E. Fix and J. L. Hodges, "Discriminatory analysis: Nonparametric discrimination," USAF School of Aviation Medicine, Randolph Field, TX, USA, Project 21-49-004, Rep. 4, 1951.

[24] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.

[25] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1007/BF00116251.

[26] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Wadsworth International Group, 1984.

[27] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," *J. Mach. Learn. Res.*, vol. 20, no. 177, pp. 1–81, 2019.

[28] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010, doi: 10.1093/bioinformatics/btq134.

[29] M. Azad, M. F. K. Khan, and S. Abd El-Ghany, "XAI-enhanced machine learning for obesity risk classification: A stacking approach with LIME explanations," *IEEE Access*, vol. 13, pp. 13847–13865, 2025, doi: 10.1109/ACCESS.2025.3530840.

[30] L. Breiman, "Statistical modeling: The two cultures," *Stat. Sci.*, vol. 16, no. 3, pp. 199–231, 2001, doi: 10.1214/ss/1009213726.

[31] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[32] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[33] S. Biswas, M. M. Islam, N. Islam, and M. A. R. Mia, "National data meets AI: Machine learning for predicting overweight/obesity among ever-married Bangladeshi women," *PLoS ONE*, vol. 21, no. 2, p. e0341821, 2026, doi: 10.1371/journal.pone.0341821.

[34] V. Gandhi *et al.*, "Benchmarking machine learning models for obesity prediction," Springer, 2025.

[35] S. Shafi *et al.*, "Optimizing machine learning models for obesity risk prediction using bio-inspired feature selection," *ScienceDirect*, 2026.

[36] W. M. Shaban *et al.*, "Machine learning framework for predicting susceptibility to obesity," *Sci. Rep.*, 2025.

[37] A. Zeeshan *et al.*, "Obesity prediction using deep neural networks on dietary data," MDPI, 2025.

[38] T. Khater *et al.*, "Explainable artificial intelligence for obesity prediction using lifestyle factors," *ScienceDirect*, 2024.

[39] Y. Görmez *et al.*, "Prediction of obesity levels based on physical activity and eating habits with a machine learning model integrated with explainable artificial intelligence," *Front. Physiol.*, vol. 16, p. 1549306, 2025, doi: 10.3389/fphys.2025.1549306.

[40] J. Teke *et al.*, "Artificial intelligence for obesity management: A review," *ScienceDirect*, 2025.

[41] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2025. [Online]. Available: https://www.R-project.org/

[42] M. Kuhn, *caret: Classification and Regression Training*, R package, 2024. [Online]. Available: https://CRAN.R-project.org/package=caret

[43] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[44] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc Functions of the Department of Statistics, Probability Theory Group*, R package, 2024. [Online]. Available: https://CRAN.R-project.org/package=e1071

[45] T. M. Therneau and B. Atkinson, *rpart: Recursive Partitioning and Regression Trees*, R package, 2023. [Online]. Available: https://CRAN.R-project.org/package=rpart

[46] X. Robin *et al.*, "pROC: An open-source package for R and S+ to analyze and compare ROC curves," *BMC Bioinformatics*, vol. 12, no. 77, 2011, doi: 10.1186/1471-2105-12-77.

[47] T. L. Pedersen and A. Benesty, *lime: Local Interpretable Model-Agnostic Explanations*, R package, 2024. [Online]. Available: https://CRAN.R-project.org/package=lime

[48] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. New York, NY, USA: Springer, 2016, doi: 10.1007/978-3-319-24277-4.

[49] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, 2nd ed. New York, NY, USA: Springer, 2021, doi: 10.1007/978-1-0716-1418-1.

[50] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Wiley, 2013.

[51] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.

[52] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[53] C. Molnar, *Interpretable Machine Learning*, 3rd ed., 2024. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[54] M. M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," *Malawi Med. J.*, vol. 24, no. 3, pp. 69–71, 2012.

[55] Y. Wu, D. Li, and S. H. Vermund, "Advantages and limitations of the body mass index (BMI) to assess adult obesity," *Int. J. Environ. Res. Public Health*, vol. 21, no. 6, p. 757, 2024.

[56] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 4765–4774.

IEER
Institute for Excellence in Education & Research