Journal of Applied Data Sciences
Vol. 7, No. 3, September 2026, pp. 2020-2034

ISSN 2723-6471
2020

Beyond Positive and Negative: A Directional SHAP Framework for Sequential Multi-Class Classification

Ahmet Yalcin[1,*], Selim Cetin[2], Bekir Cetintav[3]

[1]*Institute of Natural Sciences, Burdur Mehmet Akif Ersoy University, 15030, Burdur, Türkiye*

[2]*Department of Mathematics, Faculty of Arts and Sciences, Burdur Mehmet Akif Ersoy University, 15030, Burdur, Türkiye*

[2]*Department of Data Science and Analytics, Faculty of Arts and Sciences, Burdur Mehmet Akif Ersoy University, 15030, Burdur, Türkiye*

(Received: February 5, 2026; Revised: April 1, 2026; Accepted: June 22, 2026; Available online: July 20, 2026)

Abstract

In this study, we address the interpretative limitations of the standard Shapley Additive Explanations (SHAP) method in sequential multi-class classification problems within the scope of Explainable Artificial Intelligence (XAI). Stemming from the observation that classical SHAP is restricted to revealing only positive and negative contributions for a single class, we propose a novel directional framework that categorizes feature effects as 'Lower' (driving towards a lower class), 'Upper' (driving towards a higher class), and 'Ambiguous' (representing inconsistent effects). To validate this approach, a Random Forest model predicting obesity levels across seven hierarchical classes was trained on an open-source dataset, achieving a classification accuracy of 95.5%. Furthermore, a stability analysis comprising 10,000 sampling iterations demonstrated the robustness of the proposed framework, with dominant features retaining their directional categorizations consistently in over 99.9% of the trials. The findings indicate that unlike standard SHAP, our method successfully isolates the specific variables that prevent an instance from ascending to a higher class or descending to a lower one, particularly clarifying the role of ambiguous boundary features. In conclusion, this modification significantly enhances model transparency for complex hierarchical scenarios, and the framework has been released as an open-source Python library to provide researchers with a practical tool for automated directional feature analysis.

Keywords: Explainable Artificial İntelligence, Machine Learning, Model Explainability, Multi-Class Classification, SHAP

1. Introduction

Artificial Intelligence (AI) and Machine Learning (ML) technologies are widely utilized to address complex problems across various domains such as healthcare, finance, and engineering [1]. While the high predictive performance offered by these technologies renders them indispensable tools, the 'black-box' nature of complex models raises serious concerns regarding reliability and transparency [2], [3]. As highlighted in comprehensive surveys on black-box models, the opacity of these systems ranging from deep neural networks to ensemble methods, necessitates robust explanation methods to ensure their safe deployment [4]. The inability to comprehend how a model arrives at a specific decision hinders the adoption of these systems in critical applications and challenges the trust placed in them. Consequently, regulatory frameworks and responsible AI practices have established the explainability of model decisions as a fundamental requirement [5], [6]. In this context, Explainable Artificial Intelligence (XAI) has emerged as a prominent research field aimed at presenting the internal workings, decision-making processes, and factors influencing model outputs in a manner understandable to humans [7]. This field encompasses a taxonomy of techniques designed to produce transparent, interpretable, and socially responsible AI systems [8].

XAI methods focus not only on what the model predicts but also on why it makes such predictions, thereby aiming to enhance the auditability, fairness, and reliability of the model [2]. This is particularly crucial in the medical domain, where "medical XAI" plays a vital role in bridging the gap between advanced algorithms and clinical decision-making, ensuring that AI tools act as reliable assistants rather than obscure oracles [9]. Through these methods, model errors can be detected more easily, potential biases can be analyzed, and consequently, the societal acceptance of artificial intelligence systems is facilitated [1]. Recent systematic reviews in healthcare demonstrate that the application of XAI has grown significantly over the last decade, underscoring its necessity for validating clinical predictions [10].

***Corresponding author: Ahmet Yalcin (ahmtylcinn15@gmail.com)**
DOI: https://doi.org/10.47738/jads.v7i3.1375

ISSN 2723-6471
2021

Among the various XAI techniques developed to facilitate this transparency, the SHAP (Shapley Additive exPlanations) method stands out as one of the most prevalent and effective approaches due to its solid foundation in cooperative game theory. Although SHAP, by its very nature, performs exceptionally well in providing clear and bidirectional explanations for binary prediction tasks, certain interpretability limitations arise when applied to more complex scenarios. We focused on SHAP limitations in multi-class classification scenarios. Standard SHAP implementations are generally limited to calculating contribution values separately for each class. This limitation often proves insufficient in providing adequate insight, particularly for examples located at class boundaries or where class transitions are complex. While explaining why a model assigned an observation to a specific class, standard approaches may fail to clearly reveal the underlying dynamics that 'obscure' this assignment or hinder the transition to a higher class, or conversely, prevent it from falling into a lower class.

In this study, addressing the interpretative limitations encountered in the SHAP method within multi-class classification problems, an alternative framework has been developed. The primary objective of this study is propose a structural approach that expands its interpretative capacity. The developed framework advances beyond classical SHAP analysis; it aims not only to demonstrate the positive and negative contributions of variables in classifying an individual but also to elucidate the roles of variables that prevent the individual from upgrading to a higher class, hinder them from falling to a lower class, or fail to exhibit a clear distinction. This perspective aligns with the concept of counterfactual explanations, which seek to identify minimal changes required to alter a prediction's outcome, but adapts it to explicitly interpret directional contributions in a static SHAP framework [11]. Through the proposed modification, features are categorized into three distinct groups: 'Lower' (directing towards a lower class), 'Upper' (directing towards a higher class), and 'Ambiguous' (exhibiting conflicting contributions between classes). Consequently, this categorization endeavors to render the model's decision-making process significantly more transparent.

The primary objective of this research is to scientifically demonstrate the extent to which the proposed approach enhances model explainability and whether it provides the capability for more precise interpretations, particularly regarding class transitions. The scope of the study encompasses the development of a novel approach to the classical SHAP method and the testing of its efficacy. To this end, the proposed method was applied to a classification problem utilizing an open-access obesity dataset; the obtained results were analyzed in comparison with standard SHAP outputs, and the validity of the approach was evaluated. The significance of this study derives from the potential of this modification to the classical SHAP method to offer a more robust and reliable analytical framework in domains requiring high transparency, such as clinical diagnosis, financial risk analysis, or customer behavior prediction. Consequently, it becomes possible to gain a deeper understanding of not only what AI systems predict but also why they make such predictions, thereby enhancing trust in the decision-making mechanisms of these systems. Therefore, broadening the scope of interpretation and developing new perspectives in the field of artificial intelligence, where the need for explainability is growing day by day, are considered key factors underscoring the importance of this work.

2. Literature Review

2.1. Model-Agnostic Explanation Methods

Various explanation methods that operate independently of the model's internal structure (model-agnostic) have been developed in the literature. LIME (Local Interpretable Model-agnostic Explanations), which aims to explain the predictions of a complex model by learning a locally interpretable and faithful model in the immediate vicinity of the relevant prediction [12]. Another method, Anchors, was introduced to overcome the scope ambiguity of linear approaches like LIME, presenting high-precision 'if-then' rules (anchors) that represent 'sufficient' conditions for predictions [13]. Additionally, for deep learning models, gradient-based methods such as Grad-CAM have been proposed to provide visual explanations by highlighting important regions in the input data [14]. Although LIME and Anchors provide effective local explanations, they lack the intrinsic mechanisms required to capture directional transitions between hierarchical classes and to interpret multi-class models. SHAP's ability to measure the marginal contributions of features provides a more robust foundation for multi-class scenarios and makes it the most suitable reference point for the directed framework proposed in this study.

2.2. Calculating SHAP Values

Developed with inspiration from cooperative game theory, SHAP has garnered significant attention due to its ability to measure the contribution of each feature to the model output in a consistent and fair manner [15]. SHAP aims to quantitatively measure the contribution of each input feature to the model's prediction [15]. By referencing the model's prediction, the method calculates both the magnitude and the direction of each feature's contribution, thereby enabling

 ISSN 2723-6471
2022

the interpretation of the trained model's results. This approach is successfully employed to explain model behaviors both locally (for a single prediction) and globally (for the entire dataset). The method ensures a fair distribution of the total model output among the constituent features by computing their Shapley values.

2.2.1 Local SHAP Values

Mathematically, the Shapley value $\phi_j(i)$ for a specific feature i represents the weighted average of its marginal contributions across all possible coalitions of features [16]. This calculation is formally expressed as shown in Equation (1):

$$\phi(i) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left(\mathrm{f_x}(\mathrm{S} \cup \{\mathrm{i}\}) - \mathrm{f_x}(\mathrm{S})\right) \quad (1)$$

N represents the set of all input features. S is a subset of features excluding feature i. $\mathrm{f_x}(\mathrm{S})$ denotes the prediction of the model given the subset of features S. The term $\frac{|\mathrm{S}|!(|\mathrm{N}|-|\mathrm{S}|-1)!}{|\mathrm{N}|!}$ acts as a weighting factor, accounting for the number of permutations in which the subset S can be formed. Through this formulation, SHAP provides a consistent and locally accurate explanation for individual predictions, allowing for the decomposition of complex model behaviors into understandable feature importances.

2.2.2 Global SHAP Values

To calculate the global Shapley value, all possible coalitions of feature values must be evaluated both with and without the inclusion of the j-th feature. However, as the number of features increases, the number of possible coalitions grows exponentially, rendering the exact solution computationally infeasible for datasets with high dimensionality. To address this computational challenge, Strumbelj and Kononenko [17] proposed an approximation method based on Monte-Carlo sampling:

$$\hat{\phi}_\mathrm{j} = \frac{1}{\mathrm{M}} \sum_{\mathrm{m}=1}^{\mathrm{M}} \left(\hat{\mathrm{f}}\left(\mathrm{x}_{+\mathrm{j}}^{\mathrm{m}}\right) - \hat{\mathrm{f}}\left(\mathrm{x}_{-\mathrm{j}}^{\mathrm{m}}\right) \right) \quad (2)$$

In Equation (2), M denotes the number of iterations, x represents the instance of interest, j is the feature index, and $\hat{f}$ refers to the prediction function of the machine learning model. To resolve ambiguity, the terms $\hat{\mathrm{f}}\left(\mathrm{x}_{+\mathrm{j}}^{\mathrm{m}}\right)$ and $\hat{\mathrm{f}}\left(\mathrm{x}_{-\mathrm{j}}^{\mathrm{m}}\right)$ represent the model's predictions during the m-th iteration. Specifically, $\mathrm{x}_{+\mathrm{j}}^{\mathrm{m}}$ is a newly constructed instance where the value of feature j is taken directly from the original instance x, while the values of the remaining random subset of features are substituted with values randomly sampled from the background dataset. Conversely, $\mathrm{x}_{-\mathrm{j}}^{\mathrm{m}}$ is constructed identically to $\mathrm{x}_{+\mathrm{j}}^{\mathrm{m}}$, except that the value for feature j is also substituted with a randomly sampled value from the background dataset, rather than being taken from x. To obtain the complete set of Shapley values, this procedure must be iterated for each individual feature. This approach defines the contribution of each feature to the model based on the marginal benefit it provides in the absence of other features.

2.3. The limitation of the SHAP method in sequential multi-class classification scenarios

Recent literature highlights significant theoretical and methodological limitations of SHAP [18]. Studies have pointed out potential mathematical flaws in feature importance attribution depending on data distribution [19] and demonstrated that explanations are heavily dependent on the specific formulation of the underlying cooperative game [20]. These constraints become particularly pronounced when applied to complex classification tasks [21]. As a consequence of its underlying operational mechanism, the standard SHAP method functions efficiently in binary classification scenarios. In these cases, machine learning models predict whether a given input belongs to a target class (e.g., Class 1) or an alternative class (e.g., Class 0). Within this binary framework, SHAP allows for a clear determination of which features are influential, alongside their respective positive and negative impacts. Features with positive SHAP values are interpreted as factors driving the instance toward the target class, while simultaneously acting as barriers preventing assignment to the alternative class. However, such straightforward interpretation is not feasible in multi-class classification scenarios, particularly those involving a hierarchical or sequential relationship between classes. By examining standard SHAP values, it is impossible to definitively identify which features prevent an instance from ascending to a "higher" class or hinder it from descending to a "lower" class. This limitation stems from the fact that classical SHAP evaluates contributions solely based on a binary "presence versus absence" context within a single specific class. Consequently, this impedes the ability to perform comparative interpretations across the class hierarchy.

 ISSN 2723-6471
2023

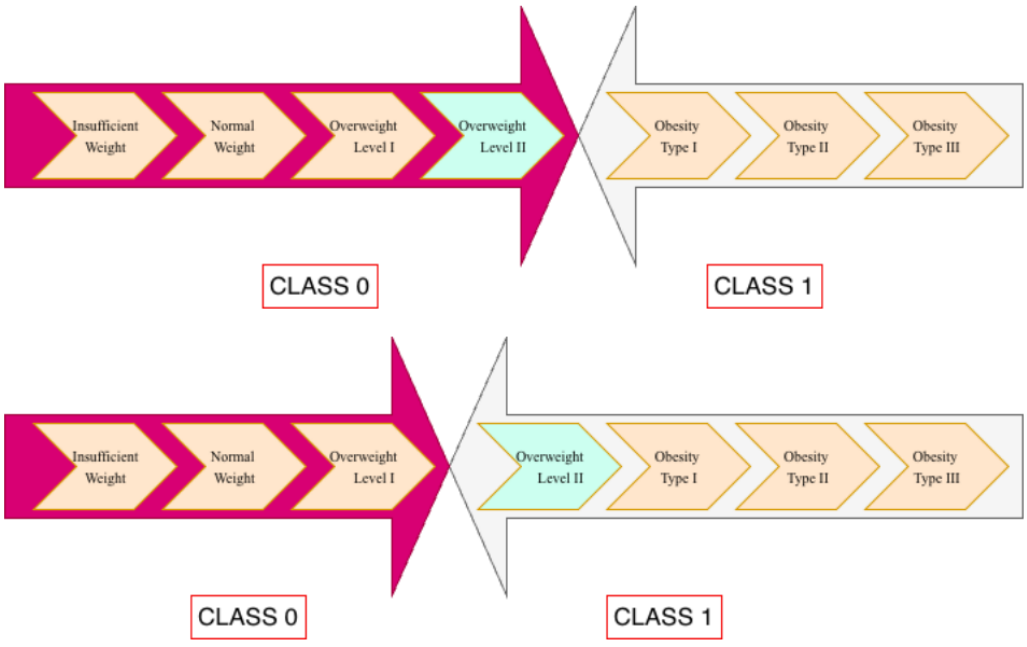
To illustrate this issue, consider the obesity classification hierarchy utilized in this study, which consists of seven sequential classes: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. Suppose the machine learning model assigns a specific individual to the "Overweight Level II" class. While classical SHAP can identify which features were effective in assigning the individual to the "Overweight Level II" class and determine their positive or negative contributions, these results yield information restricted solely to that specific class. Consequently, it is impossible to interpret whether these contributions saved the instance from descending to "Overweight Level I" or prevented it from ascending to "Obesity Type I". In other words, for this specific input, the standard method cannot provide insights into inter-class relationships or directional transitions. This situation imposes significant interpretability constraints, preventing the more profound analysis required for high-stakes domains such as clinical diagnosis.

3. Methodology

A new framework is proposed to enhance the interpretability of the SHAP method within sequential and multi-class classification scenarios. It proposes a directional framework built upon existing Shapley foundations. The primary focus is expanding its interpretative capacity to provide users with a meaningful, directional perspective on inter-class transitions. This method has been developed by leveraging the interpretability and usability that classical SHAP provides in binary classification contexts. Following this logic, multi-class structures are transformed into a 'binary class' framework. Fundamentally, this constitutes a two-stage approach (as illustrated in figure 1). The justification for utilizing precisely two stages stems from the core objective of directional analysis: to distinctly isolate the features that either prevent an instance from ascending to a "higher" class or hinder it from descending to a "lower" class. Alternative formulations, such as standard One-vs-Rest (OvR) approaches, fail to preserve the ordinal hierarchy and cannot simultaneously evaluate both upward and downward pressures. Therefore, a two-stage binary transformation is both sufficient and optimal for capturing these dual directional dynamics independently.:

First Stage: The class predicted by the machine learning model and all classes ranking *below* it are grouped together as a single class (e.g., Class 0), while the classes ranking *above* are labeled as a separate class (e.g., Class 1). SHAP values are then calculated for the input within this newly constructed binary structure. These values are recorded, and the process proceeds to the next step

Second Stage: The class predicted by the model and all classes ranking *above* it are grouped as a single class (e.g., Class 1), while the classes ranking *below* are labeled as a separate class (e.g., Class 0). SHAP values for this structure are similarly calculated and recorded.



Note: The predicted class (Overweight Level II) and all preceding hierarchical levels are consolidated into Class 0, while all subsequent higher-risk categories are grouped into Class 1.

Figure 1. Implementation of the first stage of the proposed approach.

 ISSN 2723-6471
2024

Thus, through the proposed approach, the original multi-class problem is converted into a binary classification problem in both stages. By analyzing the results obtained from these two stages, it is determined which features are effective in propelling the input towards a higher class and which are influential in pulling it towards a lower class.

To ensure complete reproducibility of the framework without strictly relying on visual references (figure 2), the underlying algorithmic logic is formally detailed as follows. Let $Y = \{y_1, y_2, \dots, y_n\}$ represent the set of sequentially ordered classes in the dataset, where n is the total number of classes. Let x denote the specific class assigned to the instance under evaluation.

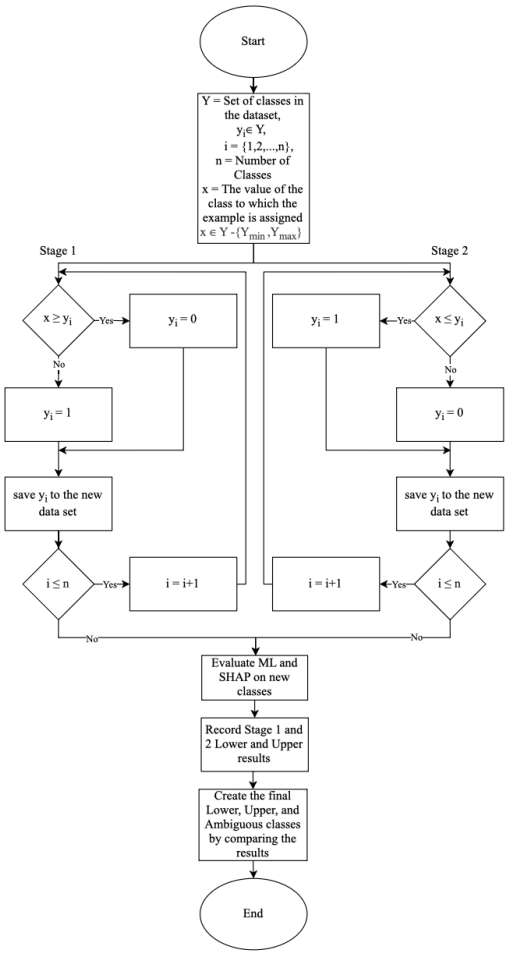


Figure 2. Flowchart Showing the Algorithm of the New Framework

The algorithm employs an iterative process $i = \{1, 2, \dots, n\}$ to relabel the dataset for both Stage 1 and Stage 2 independently. In the first stage, for each class y_i, if the condition $x \geq y_i$, is met, the new label is assigned as 0; otherwise, it is assigned as 1. Following this relabeling, the model is retrained and SHAP values are computed to record the initial 'Lower' and 'Upper' contributions. In the second stage, the condition $x \leq y_i$, is evaluated; if true, the label is assigned as 1, otherwise as 0. After a second retraining and SHAP calculation, the results from both stages are compared. Features with consistent directional impacts are assigned to the final 'Lower' or 'Upper' categories, while those yielding contradictory results across the two stages are designated as ‘Ambiguous'. This systematic comparison ensures that the directional influence of each feature is captured relative to the entire class hierarchy.

3.1. Algorithmic Procedure and Labeling Logic

In other words, in the first stage if the value of the class assigned to the instance is greater than or equal to the value of the class under consideration, it is labeled as 0; otherwise, it is labeled as 1. Based on these new labels, the Machine Learning model is retrained, and the SHAP values are computed. In the second stage, if the value of the assigned class is less than or equal to the class value under consideration, it is labeled as 1; otherwise, it is labeled as 0. The subsequent calculation steps are then repeated. In both stages, the resulting 'Lower' and 'Upper' contributions are recorded. By

ISSN 2723-6471
2025

comparing these results, the final 'Lower', 'Upper', and 'Ambiguous' values are determined, thereby concluding the analysis.

Boundary Conditions: As a natural consequence of this methodology, the application of this approach is unnecessary for inputs predicted to be in the boundary classes. In other words, the classical SHAP method suffices for the lowest and highest classes. This is because the concept of 'descending' is irrelevant for an instance in the lowest class, just as 'ascending' is inapplicable for an instance in the highest class.

3.2. Open-Source Software Implementation

To bridge the gap between theoretical advancements in Explainable Artificial Intelligence and practical usability, the directional SHAP framework proposed in this study has been engineered into a comprehensive, open-source Python library. A frequent limitation in methodological research is the lack of accessible tools; therefore, this software was specifically developed to provide researchers and data scientists with a robust, plug-and-play solution. The package fully automates the complex two-stage binary transformation and dynamically categorizes feature impacts into 'Lower', 'Upper', and 'Ambiguous' classes. Beyond backend calculations, the library features built-in visualization modules designed to generate highly intuitive, color-coded directional SHAP plots, directly addressing the interpretability constraints of standard multi-class XAI methods.

3.2.1. Technical Dependencies and Architecture

The development of the sequential-shap-explainer package heavily relies on a robust stack of open-source Python libraries. The core calculation engine leverages the standard shap library (e.g., v0.41.0 or higher) for the initial Shapley value estimations, which are subsequently processed by our custom directional algorithms. High-dimensional data manipulation and state management are efficiently handled via Pandas (v1.3.0+) and NumPy (v1.21.0+). Furthermore, the built-in visualization modules, which are responsible for rendering the detailed directional bar charts, are constructed using Matplotlib (v3.4.0+). Ensuring compatibility with these foundational libraries allows the package to integrate flawlessly into existing scikit-learn-based machine learning workflows. To ensure maximum transparency, reproducibility, and community adoption, the complete source code, detailed documentation, and practical application examples are publicly available. You can easily access the framework and integrate it into your machine learning pipelines (via PyPI using pip install sequential-shap-explainer). Available at: https://github.com/ahmtylcinn/sequential-shap.git. By open sourcing this analytical framework, this study not only introduces a novel theoretical approach but also equips the global research community with a scalable, accessible software tool tailored for complex sequential classification tasks.

3.3. Application

To validate the proposed approach, an open-access dataset was utilized. This dataset, curated by Mendoza Palechor and De la Hoz Manotas [22], is designed to predict obesity levels based on individuals' dietary habits and physical conditions. The dataset encompasses data collected from individuals aged between 14 and 61 from Colombia, Peru, and Mexico. In addition to demographic characteristics (age, gender, height, weight), the dataset includes numerous attributes reflecting daily lifestyle and dietary habits. These attributes comprise the frequency of high-calorie food consumption, vegetable consumption, number of main daily meals, consumption of food between meals, daily water intake, alcohol consumption, physical activity level, calorie monitoring, duration of technology usage, and preferred mode of transportation.

The dataset consists of a total of 2111 observations and 17 attributes (table 1). A portion of these inputs was obtained directly from participants via an online survey, while the remaining portion was synthetically generated using the SMOTE (Synthetic Minority Over-sampling Technique) method to address unbalanced class distributions. During this process, missing values and outliers were cleaned, data were normalized, and a reliable analytical environment was established. To establish a reliable analytical environment and ensure reproducibility, systematic preprocessing steps were implemented. Specifically, missing values were addressed via median and mode imputation, while outliers in continuous features were identified and treated using the standard Interquartile Range (IQR) criterion. Categorical attributes were appropriately encoded.

ISSN 2723-6471
2026

Table 1. Variables and their descriptions in the data set used

Variable Name	Variable Type	Variable Descriptions
Gender	Categorical	Gender of the individual
Age	Continuous	Age of the individual
Height	Continuous	Height of the individual
Weight	Continuous	Weight of the individual
family_history_with_overweight	Categorical	Presence of overweight family members
FAVC	Categorical	Frequency of high-calorie food consumption
FCVC	Continuous	Frequency of vegetable consumption
NCP	Continuous	Number of main meals per day
CAEC	Categorical	Consumption of food between meals
SMOKE	Categorical	Smoking habit
CH2O	Continuous	Daily water consumption
SCC	Categorical	Calories consumption monitoring
FAF	Continuous	Frequency of physical activity
TUE	Continuous	Time spent using technology devices
CALC	Categorical	Frequency of alcohol consumption
MTRANS	Categorical	Transportation used
Target	Categorical	Obesity Level (Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, Obesity Type III)

All observations were labeled in accordance with Body Mass Index (BMI) criteria, calculated based on World Health Organization (WHO) and Mexican norms. Accordingly, individuals were categorized into seven distinct classes: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. Given these characteristics, the dataset possesses a structure highly suitable for machine learning and data mining applications, particularly for classification and prediction tasks. In this study, this dataset was employed to evaluate the performance of the proposed method, to conduct comparisons with traditional models, and to test explainable artificial intelligence methods.

4. Results and Discussion

In order to evaluate the effectiveness of the proposed approach, analyses were conducted on the open-access obesity dataset [22]. The dataset comprises 2111 observations and 17 attributes reflecting individuals' demographic information, dietary habits, and physical activities. Obesity level (Insufficient Weight, Normal Weight, Overweight Level I-II, and Obesity Type I-III) was utilized as the target variable.

Initially, necessary pre-processing steps were executed on the dataset, the Random Forest algorithm was applied, and a classification accuracy of 95.5% (0.955) was achieved. This machine learning model serves purely as a demonstrative vehicle for the proposed SHAP framework. Given that the trained Random Forest model is tree-based, the Tree SHAP method was employed for explainability. Subsequently, a random sample was selected from the dataset, and its SHAP results were analyzed. The model predicted the selected instance as Overweight II, and the data values corresponding to the other variables are presented in table 2.

Table 2. The data possessed by the sample under consideration

Variable	Value
Gender	1.0 (Male)
Age	18.0
Height	1.7
Weight	78.0
family_history_with_overweight	0.0 (No)
FAVC	1.0 (Yes)
FCVC	1.0
NCP	3.0

 ISSN 2723-6471
2027

CAEC	2.0 (Sometimes)
SMOKE	0.0 (No)
CH2O	2.0
SCC	0.0 (No)
FAF	0.0
TUE	1.0
CALC	2.0 (Often)
MTRANS	3.0 (motorcycle)
Target	Overweight II

The classical SHAP method was first applied to this example. Shapley values were visualized using a waterfall graph, as shown in figure 3. Upon examination of the classical SHAP results, the features with the most significant positive and negative contributions to assigning the selected instance to the Overweight II class can be readily observed. According to the analysis, Weight (+0.26) emerges as the most dominant factor, exerting a strong influence on the individual's assignment to this class. Height (+0.09) and frequency of vegetable consumption (FCVC, +0.08) are also variables providing additional positive contributions to the classification. Furthermore, gender (being Male, +0.06), high technology usage time (TUE, +0.03), and lack of physical activity (FAF, +0.03) played a supportive role in placing the individual in the overweight category. On the other hand, factors such as daily water consumption (CH2O, −0.03) and young age (Age, −0.01) exhibited a limited mitigating effect (negative contribution) on the individual's assignment to this class. While the Classical SHAP method allows for the detection of variables contributing positively and negatively, it is not possible to interpret which variables hindered the individual from ascending to a higher class or which prevented them from descending to a lower class.

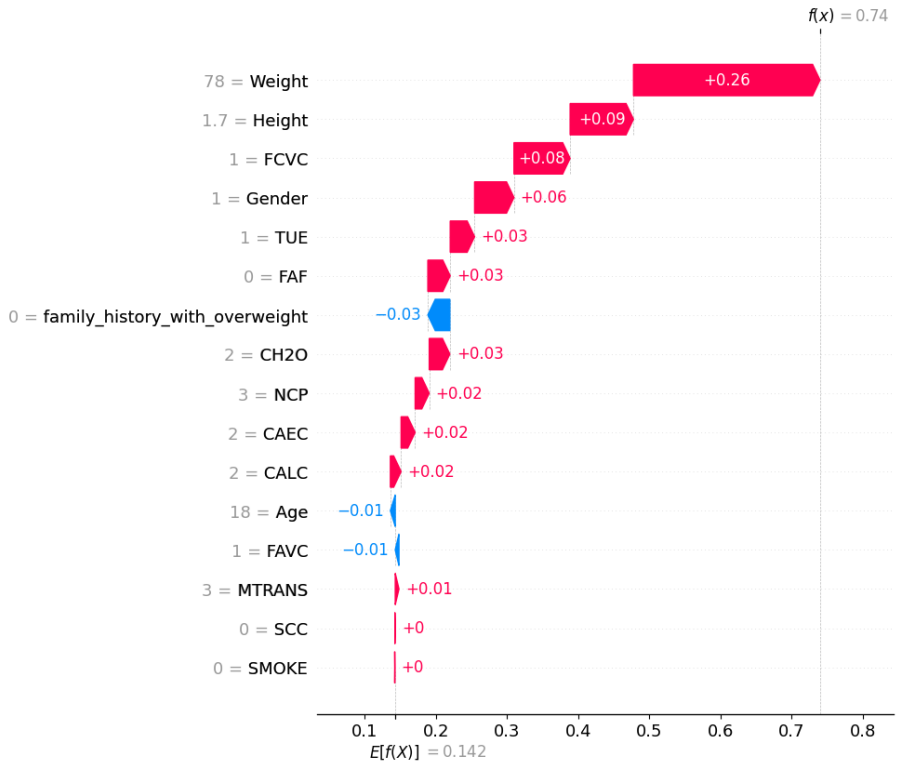


Figure 3. Classic SHAP Results for the Sample Taken

Subsequently, the first stage of the proposed solution was initiated. In this phase, the Overweight II class, to which the individual belongs, along with all classes ranking below it (Insufficient Weight, Normal Weight, Overweight I, and Overweight II), were grouped and labeled as a single class (Class 0). Consequently, Class 1 comprises the Obesity Type I, Obesity Type II, and Obesity Type III classes. Based on these class labels, SHAP values were computed and visualized using a waterfall plot, as depicted in figure 4 (A).

ISSN 2723-6471
2028

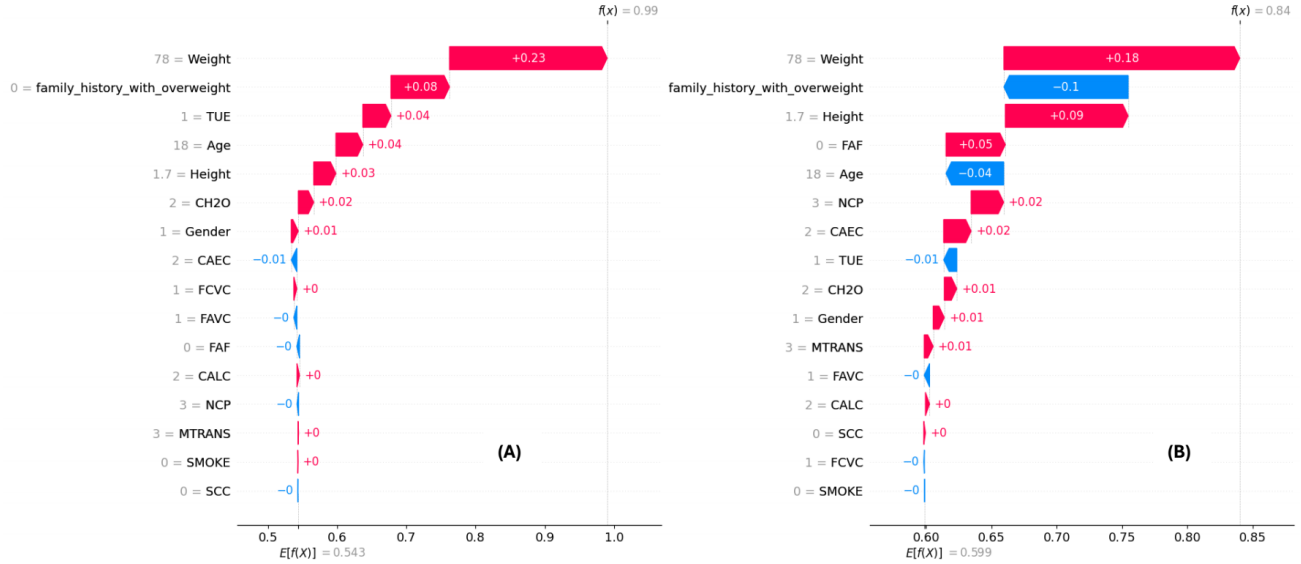


Figure 4. SHAP results for the example taken in (A) first stage, and (B) second stage of the new framework

Features exhibiting a positive contribution were recorded as 'Lower' (attributes driving the instance towards the lower class) for the initial stage; whereas features exhibiting a negative contribution were recorded as 'Upper' (attributes driving the instance towards the higher class). According to this graph, the attributes belonging to the 'Lower' and 'Upper' categories for the first stage are presented in table 3.

Table 3. Attributes belonging to the Lower and Upper Classes in the First Stage

LOWER	UPPER
Weight	CAEC
family_history_with_overweight	FAVC
TUE	FAF
Age	NCP
Height	SCC
CH2O	
Gender	
FCVC	
CALC	
MTRANS	
SMOKE	

Following the recording of the 'Lower' and 'Upper' class outputs obtained in the first stage, the second stage was initiated. In this stage, the Overweight II class, to which the individual belongs, along with the classes ranking above it (Obesity Type I, Obesity Type II, and Obesity Type III), were grouped and labeled as a single class (Class 1). Conversely, the classes ranking below (Insufficient Weight, Normal Weight, and Overweight I) were re-labeled as Class 0, and the SHAP values were subsequently calculated. These values are visualized via a waterfall plot in figure 4 (B).

Maintaining the logic of the initial labeling process, features providing a positive contribution to the individual's inclusion in Class 1 were recorded as 'Upper' (attributes driving the instance towards the higher class); whereas features providing a negative contribution were recorded as 'Lower' (attributes driving the instance towards the lower class). According to this graph, the attributes belonging to the Lower and Upper categories for the second stage are presented in table 4.

Table 4. Attributes belonging to the Lower and Upper Classes in the Second Stage

LOWER	UPPER
family_history_with_overweight	Weight
Age	Height
TUE	FAF

 ISSN 2723-6471

FAVC	NCP
FCVC	CAEC
SMOKE	CH2O
	Gender
	MTRANS
	SCC
	CALC

After recording the 'Lower' and 'Upper' class outputs obtained in the second stage, the results presented in table 3 and table 4 were compared to derive the final outcome. The underlying logic here involves identifying features that yield consistent results based on the SHAP outputs obtained in both stages, While labeling inconsistent ones as 'Ambiguous'. For instance, if a variable was recorded as 'Lower' in the first stage (implying it pushes the individual towards a lower class) but is recorded as 'Upper' in the second stage, it leads to the conclusion that this variable possesses an 'Ambiguous' nature, as it does not definitively drive the individual towards either a lower or a higher class. As a result of this comparison, the variables categorized into the final 'Lower', 'Upper', and 'Ambiguous' classes are presented in table 5.

Table 5. Attributes belonging to the final Lower, Upper, and Ambiguous Classes

Lower	**Upper**	**Ambigious**
family_history_with_overweight	FAF	FAVC
Age	NCP	Weight
TUE	CAEC	Height
FCVC	SCC	CH2O
SMOKE		Gender
		MTRANS
		CALC

Upon examination of the obtained final results (see figure 5), the attributes driving the individual towards the 'Lower' (i.e., Overweight I) direction are identified as family_history_with_overweight, Age, TUE, FCVC, and SMOKE; whereas the attributes driving the individual towards the 'Upper' (i.e., Obesity I) direction stand out as FAF, NCP, CAEC, and SCC. The attributes classified as 'Ambiguous', that is, those providing inconsistent or indecisive contributions between classes, are FAVC, Weight, Height, CH2O, Gender, MTRANS, and CALC. In order to validate the accuracy of the solution, let us examine each variable individually.

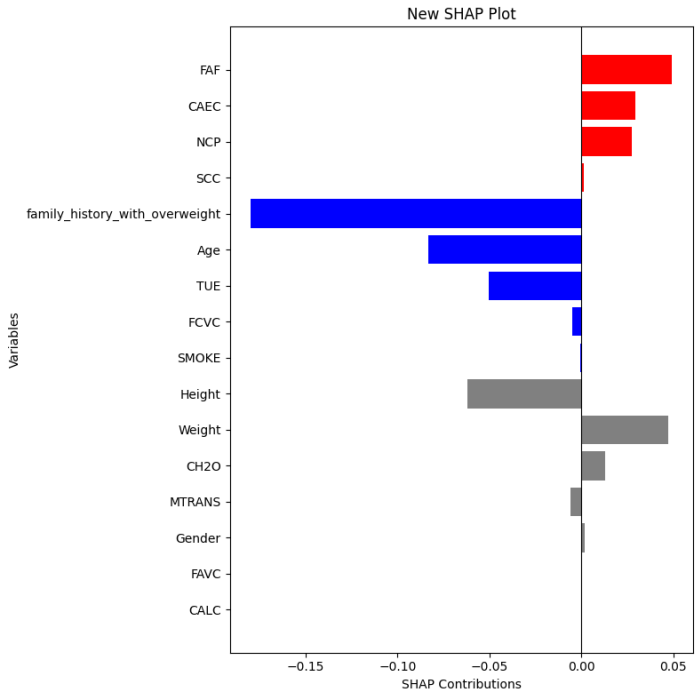


Figure 5. New SHAP Graph

 ISSN 2723-6471

Features driving the individual towards a lower obesity class (Overweight I) reveals several mitigating demographic and lifestyle factors. The absence of a family history of obesity (family_history_with_overweight = 0) and the individual's young age (18) act as primary risk-reducing factors that facilitate a tendency toward a lower obesity level. In terms of behavioral and dietary habits, a moderate duration of technology usage (TUE = 1.0) and the regular consumption of vegetables in one meal daily (FCVC = 1.0) further contribute to maintaining a healthier classification. Finally, the individual's non-smoking status (SMOKE = 0) serves as an indicator of a generally healthy lifestyle, acting synergistically with the other variables to direct the model's prediction towards a lower risk category.

The factors driving towards a higher obesity class are as follows: the lack of physical activity (FAF = 0.0) is a strong indicator that increases obesity risk, serving as a primary factor propelling the individual towards Obesity I. Furthermore, a relatively high number of daily meals (NCP = 3.0), alongside other values, increases overall energy intake and exhibits a driving effect towards this upper class. The occasional consumption of food between meals (CAEC = 2.0) is another behavioral factor that increases the tendency towards obesity. Finally, the lack of calorie monitoring (SCC = 0) diminishes the individual's nutritional awareness, providing an additional driving contribution towards the upper classification.

In addition to these variables, it can be stated that the attributes belonging to the Ambiguous class neither direct the individual towards a higher class nor prevent them from remaining in a lower class. This is because these variables did not yield consistent results across the two stages of the classification process. Consequently, it can be considered that the attributes in question indirectly contribute to the individual remaining in their current class (Overweight II).

4.1. Robustness and Stability Analysis of Directional Categorizations

To ensure the statistical reliability and stability of these local (instance-based) categorizations, a robustness test comprising 10,000 iterations of the sampling process was conducted for the selected observation. As illustrated in the boxplot (figure 6), the model exhibits a high degree of stability across the 10,000 iterations.

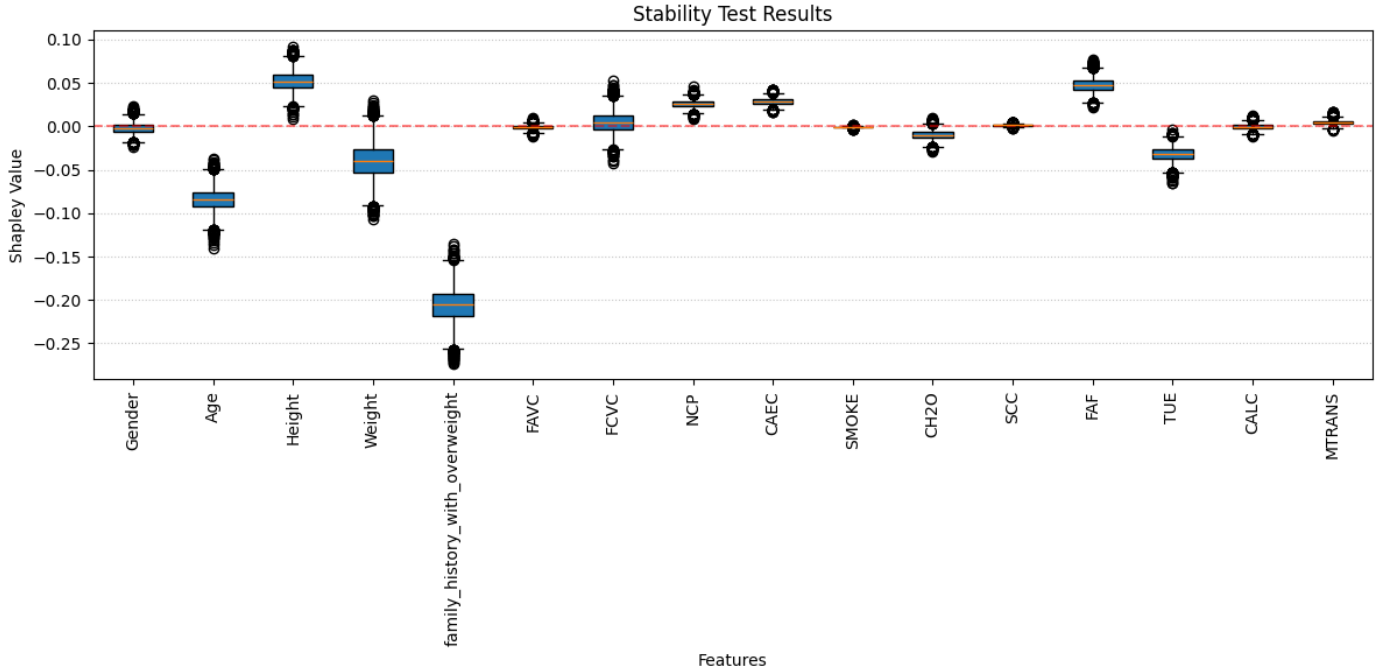


Figure 6. Boxplot illustrating the stability of Shapley values across 10,000 sampling iterations

The notably narrow interquartile ranges (tight boxes) for the vast majority of features indicate very low variance in the calculated Shapley values. This minimal spread demonstrates that the underlying directional calculations are highly consistent and robust against computational noise, ensuring reliable feature categorizations. Furthermore, table 6 details the percentage distribution of each feature's directional categorization across the 10,000 iterations, alongside their mean Shapley values and standard deviations. These numerical results strongly corroborate the high consistency of the proposed framework. Features with dominant directional impacts maintained their classifications with remarkable reliability.

ISSN 2723-6471

Table 6. Percentage distribution of directional categories and statistical summary across 10,000 iterations

Feature	Category Distribution (Upper, Lower, Ambiguous)			Mean	Std Dev
FAF	99.94	0	0.06	0.047371	0.007409
CAEC	100.0	0	0	0.028986	0.003642
NCP	85.84	0	14.16	0.025790	0.004189
SCC	80.51	0.71	18.78	0.001381	0.000801
family_history_with_overweight	0	100.0	0	-0.205563	0.019329
Age	0	100.0	0	-0.084289	0.012762
TUE	0	66.09	33.91	-0.032373	0.007789
FCVC	16.53	65.86	17.61	-0.004444	0.011698
SMOKE	0	89.83	10.17	-0.000662	0.000314
Height	0	0	100.0	0.051952	0.010736
Weight	0	0	100.0	-0.039664	0.019227
CH2O	0	14.18	85.82	-0.009651	0.004978
MTRANS	13.17	0.37	86.46	0.004641	0.002711
Gender	0	0.24	99.76	-0.001481	0.006052
FAVC	0.84	0	99.16	-0.001156	0.002291
CALC	1.52	21.8	76.68	-0.000460	0.003086

For instance, highly influential features like CAEC, family_history_with_overweight, Age, and FAF consistently retained their respective directional categories ('Upper' or 'Lower') in over 99.9% of the trials. Similarly, core 'Ambiguous' attributes (e.g., Height and Weight) demonstrated exceptional stability within their classification. While features with mean Shapley values near zero (such as FCVC and TUE) exhibited expected distributional variance, their extremely low standard deviations confirm these fluctuations are negligible. Ultimately, these findings validate that the framework provides highly reproducible categorizations reflecting stable model dynamics.

The obtained findings indicate that the interpretative capacity offered by the classical SHAP method for multi-class problems is limited; conversely, the proposed modified approach is capable of providing more detailed and meaningful explanations by taking into account the distinction between lower and upper classes. Thereby, attributes that contribute to individuals remaining in a specific class, hinder them from ascending to a higher class, or prevent them from descending to a lower class have been clearly demonstrated. It is observed that the proposed method renders the decision-making process more transparent, particularly for observations situated at class boundaries, and elucidates the effects of 'Ambiguous' attributes. Thus, not only have the performance nuances of the model been better analyzed, but a significant improvement in the level of explainability has also been achieved compared to the classical SHAP method. Consequently, it is evaluated that the developed approach can present a valid and robust solution for various multi-class classification problems possessing a hierarchical structure.

The graph in figure 5 was designed to comprehensively visualize the results of the modified SHAP approach proposed in this study. By presenting the contribution of each variable to different class orientations at a single glance, the graph provides an opportunity for detailed interpretation beyond the scope of the classical SHAP method. In this visualization process, the Shapley values obtained in the first and second stages were taken into account, and these values were summed to calculate the final magnitude of effect. Thus, the status of each variable, whether directing the individual towards a higher class ('Upper'), ensuring they remain in a lower class ('Lower'), or exhibiting unstable/conflicting effects ('Ambiguous'), has been clearly demonstrated.

In the graph, the red bars represent attributes in the 'Upper' category that contribute to the individual's transition to a higher obesity class. The most dominant variable in this category is FAF, representing a lack of physical activity, which is observed to have the strongest driving effect towards obesity. Similarly, the snacking habit CAEC and the daily meal count NCP are among other significant variables driving the individual towards the upper class. Furthermore, although the SCC variable is in the Upper class, its contribution value is relatively negligible; therefore, its effect in propelling the individual to the upper class can be considered minimal. The blue bars indicate attributes in the 'Lower' category that support the individual remaining in a lower class. In particular, the directing effects towards the lower class of factors such as the absence of family history of obesity (family_history_with_overweight), young age (Age), and non-smoking status (SMOKE) are noteworthy. These results demonstrate that the attributes in question play a restrictive role regarding the individual's obesity risk.

Finally, the grey bars reflect the variables falling into the 'Ambiguous' category. These attributes did not consistently contribute to the upper or lower class across both stages; they exhibited an unstable structure by displaying effects sometimes upwards and sometimes downwards. For instance, the placement of variables such as Weight, Height, and CH2O in the Ambiguous class suggests that these variables provide indirect contributions to the individual remaining in their current class. Consequently, figure 5 systematically differentiates the distinct effects of attributes on class transitions, transcending the interpretative capacity of the classical SHAP method, which is limited to positive and negative contributions. Thereby, a more transparent and comprehensive framework for interpretation is presented.

5. Conclusion

In this study, the limitations of the SHAP method in multi-class classification problems were addressed, and an alternative solution was proposed. The primary motivation was to overcome the inability of classical SHAP to explain the directional dynamics that prevent an instance from ascending to a higher class or descending to a lower one. Experimental results on the obesity dataset demonstrated that the proposed modification effectively categorizes the directional dynamics between hierarchical classes, offering a specific interpretative perspective that overcomes the limitations of the classical method. By categorizing features into 'Lower' (directing towards a lower class), 'Upper' (directing towards a higher class), and 'Ambiguous' (inconsistent contributions), the model's decision-making process was rendered more transparent. For instance, while factors such as *lack of physical activity* were clearly identified as driving forces towards the Upper class, variables like *weight* and *height* were revealed to have Ambiguous roles in specific contexts, a nuance often missed by standard approaches.

Consequently, this research establishes that the proposed framework offers a practical and extended interpretative perspective for hierarchical multi-class classification scenarios. Rather than introducing a competing mathematical formula, this approach successfully enriches the existing interpretative capacity of the classical SHAP method by providing a nuanced, directional understanding of sequential class transitions. A key limitation of the proposed framework is the computational cost arising from the model being completely retrained at both stages. Whilst this doubling of the training burden may be manageable for fast, tree-based models such as Random Forests, it can create scalability issues for computationally intensive algorithms such as those used with large datasets or deep neural networks.

6. Declarations

6.1. Author Contributions

Conceptualization: A.Y., B.C. and S.C.; Methodology: S. C., B.C. and A.Y.; Software: A.Y. and B.C.; Validation: A.Y. and B.C.; Formal Analysis: A.Y., B.C., and S.C.; Investigation: A.Y. and B.C.; Resources: A.Y., S.C.; Data Curation: A.Y. and B.C.; Writing Original Draft Preparation: A.Y., S.C., and B.C.; Writing Review and Editing: A.Y., S.C., and B.C.; Visualization: A.Y. and B.C.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The obesity dataset used in this study is an open-access dataset (Available at: https://www.kaggle.com/datasets/muhramasaputra/obesity-based-on-eating-habits-and-physical-cond) . The software framework developed (sequential-shap-explainer) is available via PyPI and GitHub (https://github.com/ahmtylcinn/sequential-shap.git).

6.3. Funding

This work was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) BİDEB 2211-National Graduate Scholarship Program.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing 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.

References

[1] E. Papagiannidis, P. Mikalef, and K. Conboy, "Responsible artificial intelligence governance: A review and research framework," *J. Strategic Inf. Syst.*, vol. 34, no. 2, pp. 1-18, 2025, doi: 10.1016/j.jsis.2024.101885.

[2] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," *arXiv*, vol. 2020, no. June, pp. 1-24, 2020, doi: 10.48550/arXiv.2006.11371.

[3] C. Meske, B. Abedin, M. Klier, and F. Rabhi, "Explainable and responsible artificial intelligence," *Electron. Markets*, vol. 32, no. 4, pp. 2103–2106, 2022, doi: 10.1007/s12525-022-00607-2.

[4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2018, doi: 10.1145/3236009.

[5] S. Goellner, M. Tropmann-Frick, and B. Brumen, "Responsible artificial intelligence: A structured literature review," *arXiv*, vol. 2024, no. March, pp. 1-54, 2024, doi: 10.48550/arXiv.2403.06910.

[6] Y. Wang, M. Xiong, and H. Olya, "Toward an understanding of responsible artificial intelligence practices," in *Proc. 53rd Hawaii Int. Conf. Syst. Sci. (HICSS)*, Maui, HI, USA, vol. 2020, no. Jan., pp. 4962–4971, 2020, doi: 10.24251/HICSS.2020.610.

[7] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: An analytical review," *Wiley Interdiscip. Rev. Data Mining Knowl. Discov.*, vol. 11, no. 5, pp. 1-13, 2021, doi: 10.1002/widm.1424.

[8] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, no. June, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[9] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, 2021, doi: 10.1109/TNNLS.2020.3027314.

[10] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Comput. Methods Programs Biomed.*, vol. 226, no. November, p. 107161, 2022, doi: 10.1016/j.cmpb.2022.107161.

[11] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. Conf. Fairness, Accountability, and Transparency (FAT)*, vol. 2020, no. January, pp. 607–617, 2020, doi: 10.1145/3351095.3372850.

[12] 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. Discovery Data Mining (KDD)*, San Francisco, CA, USA, vol. 2016, no. Aug., pp. 1135–1144, 2016, doi: 10.1145/2939672.2939778.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, pp. 1-9, New Orleans, LA, USA, Apr. 2018, doi: 10.1609/aaai.v32i1.11491.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2016, no. October, pp. 618–626, 2016, doi: 10.48550/arXiv.1610.02391.

[15] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst.*, vol. 30, Long Beach, CA, USA, vol. 2017, no. May, pp. 4765–4774, 2017, doi: 10.48550/arXiv.1705.07874.

[16] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Morrisville, NC, USA: Lulu Press, 2020.

[17] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowl. Inf. Syst.*, vol. 41, no. 3, pp. 647–665, 2014, doi: 10.1007/s10115-013-0679-x.

[18] S. J. Silva and C. A. Keller, "Limitations of XAI methods for process-level understanding in the atmospheric sciences," *Artif. Intell. Earth Syst.*, vol. 3, no. 1, pp. 1-6, 2024, doi: 10.1175/AIES-D-23-0045.1.

 ISSN 2723-6471

[19] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 2020, no. February, pp. 5491–5500, 2020, doi: 10.48550/arXiv.2002.11097.

[20] L. Merrick and A. Taly, "The explanation game: Explaining machine learning models using Shapley values," in *Machine Learning and Knowledge Extraction*. Cham, Switzerland: Springer, vol. 2020, no. August, pp. 17–38, 2020, doi: 10.1007/978-3-030-57321-8_2.

[21] Y. Takefuji, "Limitations of SHAP-based interpretations in environmental and membrane filtration applications," *Water Res.*, vol. 288, no. January, pp. 1-5, 2025, doi: 10.1016/j.watres.2025.124766.

[22] 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, no. August, pp. 1-5, 2019, doi: 10.1016/j.dib.2019.104344.