Obesity Prediction using Bayesian Optimized Gradient Boosted Trees

EDRIC CASTEL HAO

De La Salle University - Manila

MADELEIN VILLEGAS

De La Salle University - Manila

DINO DOMINIC LIGUTAN

De La Salle University - Manila

Obesity is a growing concern worldwide, and its prediction is crucial for preventing and managing related health problems. In this study, we aim to develop a reliable and accurate model for predicting obesity based on various demographic and lifestyle factors. Prior studies have been trying to combat obesity by understanding the underlying causes and developing interventions to treat and prevent the disease. The model considers data such as age, gender, aspects of physical activity, and dietary habits from the publicly available dataset in the University of California Irvine (UCI) Machine Learning Repository to make predictions. We used different machine learning algorithms to analyze the data and evaluate the performance of the model. Our results showed that the model had a high accuracy rate of 99.84% from the LightGBM model, which indicates its potential for practical use. Nested Stratified Cross-Validation (CV) was used to confirm the results of the model prediction. Furthermore, the model provides valuable insights into the factors that contribute to obesity, which can be used to inform public health policies and interventions. In conclusion, this study has important implications for the prevention and management of obesity and highlights the importance of considering demographic and lifestyle factors in obesity prediction.

CCS CONCEPTS • Applied Computing • Life and Medical Sciences • Health Informatics

Additional Keywords and Phrases: Obesity Prediction, Gradient Boosting, Bayesian Optimization, LightGBM, XGBoost, CatBoost, SHAP

ACM Reference Format:

Edric Castel Hao, Madelein Villegas, and Dino Dominic Ligutan. 2023. Obesity Prediction using Bayesian Optimized Gradient Boosted Trees. In ICMHI '18 International Conference on Medical and Health Informatics, May 12-14, 2023, Kyoto, Japan.

1 INTRODUCTION

Obesity has been a global problem for thousands of years; evidence of being overweight has reached back more than 20,000 years [1]. Obesity, on the other hand, was not recognized as a serious health risk until the late twentieth century. Obesity prevalence began to rise significantly in the 1980s, concerning the scientific community. Early obesity research centered on deciphering the physiological systems that govern body weight. Eventually, researchers also began to investigate the influence of obesity on health, discovering several obesity-related illnesses such as heart disease, type 2 diabetes, and some forms of cancer [2].

Since then, obesity research has been an active area of research, with a focus on understanding the causes, consequences, and treatment of this health concern. Researchers have studied the genetic, environmental, and behavioral factors that contribute to weight gain, and the impact of obesity on health. There are several lifestyle factors proven to contribute to the development of obesity such as poor diet, sedentary lifestyle, unhealthy eating habits, and poor sleep habits among others [1,2]. The interaction and reinforcement of these factors are important to note, illustrating the difficulty of addressing obesity through individual lifestyle changes alone.

As the understanding of obesity grew throughout the years, researchers began to focus on developing effective interventions for weight loss and management. This led to the testing of various diets, physical activity programs, and pharmacotherapies, as well as the development of bariatric surgery in 1954. In recent years, there has been a growing recognition of the complex, multi-factorial nature of obesity, leading to a shift towards a more holistic approach to addressing the condition. This approach considers not only individual factors, such as diet and physical activity, but also the impact of social and economic factors, such as access to healthy food and opportunities for physical activity [3].

Machine learning is increasingly being used in studies of obesity as a tool to analyze and predict this condition. Using data and algorithms, the complex and multi-factorial nature of obesity is used to develop more effective interventions for weight loss and weight management. In some cases, feature engineering and data preprocessing techniques are employed to improve the accuracy of the predictions, which is done in this study. The effectiveness of the models created is validated through various methods, such as cross-validation and other resampling techniques, to ensure their reliability. The application of machine learning to the study and prediction of obesity is assisting researchers in gaining a better knowledge of the problem and developing effective management and prevention measures.

This research aims to build upon previous efforts in enhancing the accuracy of estimating obesity. The goal of this study is to consider the numerous lifestyle variables that contribute to the development of obesity. Researchers can utilize Machine Learning to better understand the influence of specific elements that may not be obvious from aggregated data. This study applies a tree-based supervised algorithm that evaluates numerous lifestyle indicators and health information to determine whether a person is obese.To validate the findings of the study, the researchers used a resampling technique called nested stratified cross-validation to evaluate the model's effectiveness.

2 RELATED WORKS

Cervantes and Palacio calculated the percentage of overweight children aged 2 to 17 years in small geographic regions and created a logistic regression model based on survey participant variables to predict the individual's likelihood of having a high BMI. They combined this model with a simulation to estimate prevalence among subpopulations using data from various American surveys and census [4].

De-La-Hoz-Correa, Mendoza-Palechor, et. al have implemented three machine learning algorithms: Decision Trees, Naïve Bayes and Logistic Regression to estimate obesity levels in people from Mexico, Peru, and Colombia. 97.4% accuracy was achieved using the Decision Trees model, compared to Bayesian Networks and Logistic Regression [5]. This study also gave insights into the relevancy of eating habits and physical condition to a person’s overall obesity level. In [6] using the same data from [5], unsupervised data mining techniques were used to determine obesity levels based on certain lifestyle factors. This study made use of the

Simple K-Means, Decision Trees, and Support Vector Machines algorithms to determine the best algorithm to predict obesity. The resulting algorithm was a Decision Tree combined with the K-Means algorithm which resulted in 98.5% precision and recall. Gradient Boosting algorithms have also been used to predict obesity as in [7] which used the XGBoost model to predict the onset of early childhood obesity based on electronic health records from birth up to two years of age.

The most recent paper making use of the UCI ML Repository dataset involves [8], the author detecting and predicting obesity with the use of Extremely Randomized Trees, Multilayer Perceptron and XGBoost. This study accurately predicted (99.43%) obesity using a hybrid model combining the Multilayer Perceptron Model and XGBoost algorithms. Unfortunately, the previous study failed to check the results for overfitting. This research takes a comprehensive approach to feature engineering, data pre-processing, and nested stratified cross-validation, which were not adequately addressed in previous studies.

3 METHODOLOGY

3.1 Data Collection and Preprocessing

The data was sourced from the UC Irvine Machine Learning Repository. It has 2111 instances with 17 attributes namely: Gender, Age, Height, Weight, family history, frequent consumption of high caloric food (FCHCF), frequency of consumption of vegetables (FCV), number of main meals (NMM), consumption of food between meals (CFBM), Smoke, consumption of water (CW), calorie consumption monitoring (CCM), physical activity frequency per week (PAF), time using technology devices a day (TUT), consumption of alcohol (CA), Transportation, and Obesity. The dataset has some columns labeled differently which the researchers renamed for consistent labeling such as FAVC to FCHCF and FCVC to FCV. The dataset is also not stratified which often leads to imbalanced features during train-test splits.

To prepare the data for input into several types of machine learning models, the categorical features were converted into an ordinal integer datatype. For the binary categories (e.g., Male/Female, yes/no) it can be converted into 0, 1, or vice versa with no issue as the order is preserved either way. Some of the other categorical features indicate the order through the description (e.g., no/sometimes/frequently/always, Insufficient_Weight/Normal_Weight/Overweight_Level_I/Overweight_Level_II/Obesity_Level_I/Obesity_Level_II/Obesity_Level_III). However, the feature transportation does not indicate any order, so the researchers used the average obesity level for each transportation mode to rank them in this order from least to most obese: Walking, Bike, Motorbike, Automobile, and Public Transportation.

3.2 Modelling and Hyperparameter Tuning

To prepare the data for input into several types of machine learning models, the categorical features were converted into an ordinal integer datatype. For the binary categories (e.g., Male/Female, yes/no) it can be converted into 0, 1, or vice versa with no issue as the order is preserved either way. Some of the other categorical features indicate the order through the description (e.g., no/sometimes/frequently/always, Insufficient_Weight/Normal_Weight/Overweight_Level_I/Overweight_Level_II/Obesity_Level_I/Obesity_Level_II/Obesity_Level_III). However, the feature transportation does not indicate any order, so the researchers used the average obesity level for each transportation mode to rank them in this order from least to most obese: Walking, Bike, Motorbike, Automobile, and Public Transportation.

Kirk, Kok, et al. outlined guidelines for applying machine learning techniques to datasets in nutrition research especially those in obesity, diabetes, cancer, and cardiovascular diseases. In medical situations with dire consequences, interpretable models are preferred over ensemble models despite the decreased performance. However, methods such as SHAP, LIME, and partial dependence plots allow researchers to indicate the features most important for the model decision while preserving accuracy. Overfitting was mentioned as an unaccounted problem in the literature [9].

The most common validation method to combat overfitting is data splitting but there can be large variations in accuracy for any given split, so it is not sufficient to assess the generalizability of the model. Cross-validation addresses this problem by testing the model multiple times with different splits. An important variation is stratified cross-validation, which balances the proportion of the target classes to avoid cases where there are too few examples from which the model can learn. A new issue arises when researchers perform hyperparameter optimization on a single cross-validation scheme. The model is exposed to all the data during optimization even if not all at once so a certain level of overfitting still occurs. Finally, nested cross-validation addresses all these concerns by creating two layers of cross-validation. The inner layer is used to optimize the hyperparameters which are then tested on the hold-out data from the outer layer for validation. The computational cost increases substantially but apart from testing on external data this is the most honest measure of model accuracy [9]. Prior studies fail to test the effectiveness of the performance of their algorithms against overfitting and generalizability.

The results of a black box optimization challenge were described in [10] placing a huge focus on hyperparameter optimization. To combat overfitting, they use concealed machine learning models to optimize a secret evaluation metric applied to an undisclosed dataset. Aside from that participants only have access to the feedback leaderboard during the competition but will be ranked by the final leaderboard which will run their final models only once. They also limited the number of optimization iterations to better fit it to real-world model training scenarios. Out of 65 teams, 61 managed to beat the baseline random search. Moreover, all the top 20 teams used a surrogate function and most of them also used a gaussian process to model the prior distribution. This indicates that Bayesian-type of optimization methods are vastly superior to random search and by extension grid search with the top submissions reaching 100x efficiency gains. Prior studies on this dataset fail to indicate their hyperparameter optimization methods whenever relevant.

Lundberg and Lee proposed an explanation model that unifies six previous methods under the class of additive feature attribution models. These models approximate the more complex model for a specific input with linear functions of binary variables. They show that their proposed framework Shapley Additive Explanations (SHAP) provides a unique solution within that class of models with three desirable properties. The first is local accuracy which means that the output of the simplified model with the simplified input must be equivalent to the output of the original model with the original input. The second is missingness which is a property of all the previous methods. It ensures that features missing in the original model have no impact on the simplified model. The third and final property is consistency, which means that if the original model changes such that a feature's contribution does not decrease, its attributed contribution in the simplified model must also not decrease. This provides the mathematical foundation for using SHAP values to rank feature importance and remove underperforming features that merely add noise to the model [11].

On tabular datasets, Lundberg et al. find that tree-based models, particularly gradient-boosted models, frequently outperform deep learning techniques. In reality, due to model mismatch effects with complicated cancellations of irrelevant data, tree-based models can be more interpretable than basic linear models. As a result, explaining AI places a high value on deciphering tree-based models. The authors proposed TreeExplainer, which leverages the tree-based internal structure of the trained model to reduce the exact computation of Shapley values from exponential to low-order polynomial time. This allows for the calculation and modeling of Shapley values for tree-based models within a reasonable amount of time [12].

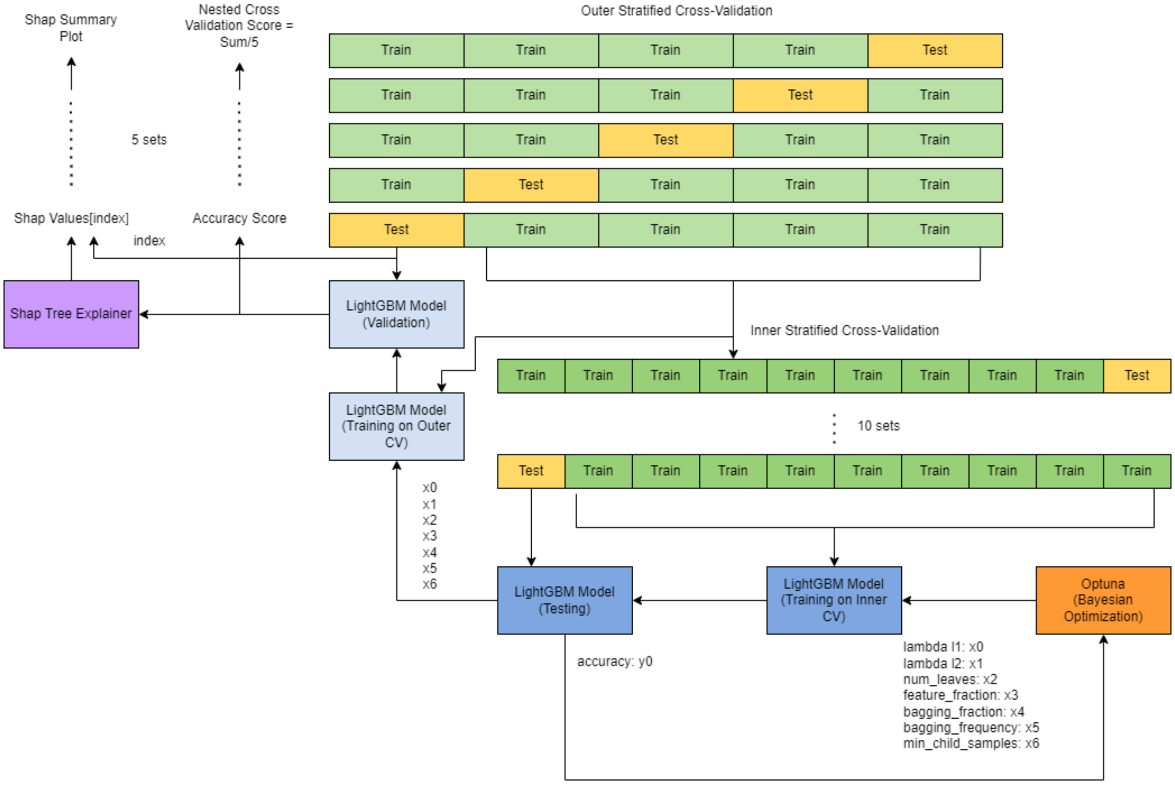
Shap Summary Plot
Nested Cross Validation Score = Sum/5
Outer Stratified Cross-Validation
Train Train Train Train Test
Train Train Train Test Train
Train Train Test Train Train
Train Test Train Train Train
Test Train Train Train Train
5 sets
Shap Values[index]
Accuracy Score
index
Shap Tree Explainer
LightGBM Model (Validation)
LightGBM Model (Training on Outer CV)
Inner Stratified Cross-Validation
Train Train Train Train Train Train Train Train Train Test
10 sets
Test Train Train Train Train Train Train Train Train Train
x0
x1
x2
x3
x4
x5
x6
LightGBM Model (Testing)
LightGBM Model (Training on Inner CV)
Optuna (Bayesian Optimization)
accuracy: y0
lambda l1: x0
lambda l2: x1
num_leaves: x2
feature_fraction: x3
bagging_fraction: x4
bagging_frequency: x5
min_child_samples: x6

Figure 1: Nested Cross-Validation Procedure

Following the best practices discussed above, the researchers decided to test multiple gradient-boosting models namely: XGBoost, LightGBM, and CatBoost starting with both a 70-30 train test split and a 5-fold cross-validation to choose the one most promising in terms of time and accuracy. As seen in figure 1, the best model will then undergo additional validation using nested stratified cross-validation. To assess the relative importance of each feature and weed out "noisy" characteristics, SHAP values calculated during validation using TreeExplainer would be employed. To identify its final best accuracy, the model will then be revalidated.

4 RESULTS

Table 1: Performance of Tested Gradient Boosting Models

Techniques	Gradient Boosting Models		
	XGBoost	LightGBM	CatBoost
70/30 Train-Test Split	99.68%	99.84%	99.68%

Techniques	Gradient Boosting Models		
	XGBoost	LightGBM	CatBoost
5-fold Stratified Cross-Validation	99.05%	99.10%	99.05%

Table 1 shows that LightGBM is the best-performing model when conducting the basic validation methods, the researchers decided to further validate its performance using nested stratified cross-validation for both the original dataset and a modified version where Overweight_level_I and Overweight_level_II are simply classed as Overweight as in some of the previous studies.

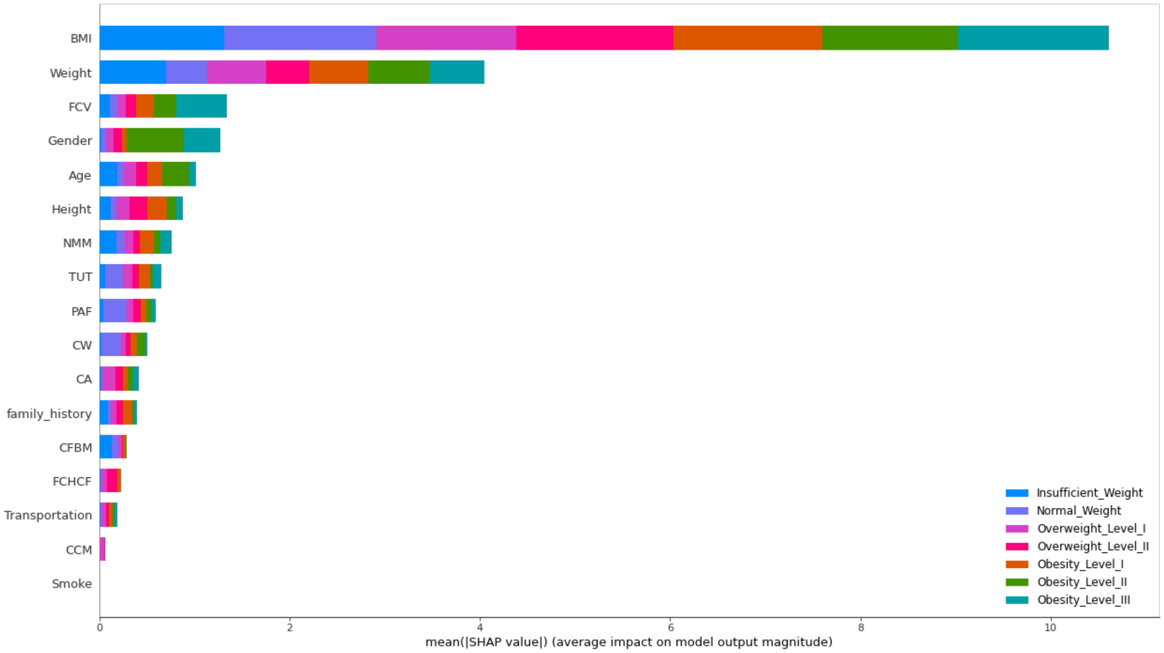

Figure 2: SHAP Summary Plot of the Original Dataset

The SHAP Summary Plots in Figure 2 indicate that BMI is unsurprisingly the most important feature as the obesity classification scale is based on this metric with some adjustments. These adjustments naturally lead to weight being the second most important feature. Likewise, gender, age, and height are common things to consider when classifying someone's obesity level. A non-obvious feature near the top of the list is FCV or frequency of consumption of vegetables. While consuming vegetables by themselves does not eliminate obesity, people who consume more vegetables also make healthier lifestyle choices hence the reduced rate of obesity. We note that FCV is more important in distinguishing between obesity levels 2 and 3 than with lower weight classes indicating that it is merely deciding at the margins.

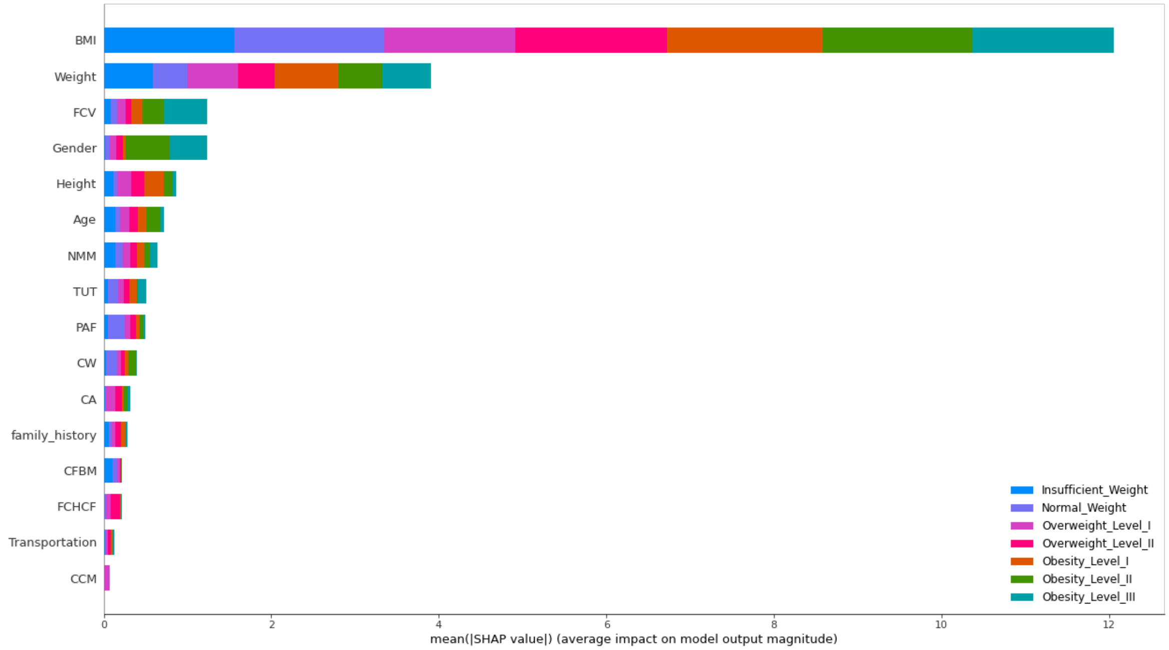

Figure 3: SHAP Summary Plot of Original Dataset with Smoke Feature Removed

As seen from Figures 2 and 3, smoke does not contribute to the model prediction and may merely be adding noise. The removal of this feature with slight accuracy gains as seen in Table 2 confirms this assessment. To provide a comparison with previous studies that only had 6 classes, the researchers also validated the modified version of the dataset.

Table 2: Validation Scores of the Three Datasets

Dataset	Average	Nested Stratified Cross-Validation Scores				
Original	98.96%	98.58%	99.29%	98.82%	99.53%	98.58%
Smoke Removed	99.01%	98.82%	99.29%	98.82%	99.76%	98.34%
Modified	99.38%	99.76%	99.29%	99.53%	99.29%	99.05%

Table 3: Optimal Hyperparameters of Proposed Models

Parameters	Optimal Values After Tuning	
	Modified dataset's best parameters	Original dataset's best parameters
lambda_l1	4.3434x10^-8	1.0802 x 10^-4
lambda_l2	0.4229	1.1455 x 10^-8
num_leaves	96	185
feature_fraction	0.4462	0.5744
bagging_fraction	0.8331	0.8942
bagging_freq	1	5
min_child_sample	21	5

It can be noted that the modified dataset required a more significant amount of L2 regularization and a reduced number of leaves to prevent overfitting as the dataset complexity is reduced by merging the two types of overweight.

5 CONCLUSION

This study identified the risk factors that contribute to overweight and obesity using machine learning from the publicly available dataset from the University of California Irvine (UCI) Machine Learning Repository. Using the prediction method of Bayesian Optimized Gradient Boosted Trees, the researchers were able to create a model with an accuracy of 99.84% which surpasses the works of previous studies using the same dataset. We also used Nested Stratified CV to validate the results of the model. There are no massive discrepancies with an accuracy of 99.01%.

By repeating the nested CV at least 100 times, future studies may try to validate the study's findings. Additionally, research must be done to provide more datasets with uniform attributes from a wide range of participants. To prevent obesity before it occurs, it is also recommended to build models that could predict obesity based solely on lifestyle characteristics.

REFERENCES

[1] G. Williams and Frühbeck Gema, *Obesity: Science to practice*. Chichester, UK: Wiley, 2009.

[2] A. Hruby, J. A. E. Manson, L. Qi, V. S. Malik, E. B. Rimm, Q. Sun, W. C. Willett, and F. B. Hu, “Determinants and consequences of obesity,” *American Journal of Public Health*, vol. 106, no. 9, pp. 1656–1662, 2016.

[3] Lee A, Cardel M, Donahoo WT. Social and Environmental Factors Influencing Obesity. [Updated 2019 Oct 12]. In: Feingold KR, Anawalt B, Blackman MR, et al., editors. Endotext [Internet]. South Dartmouth (MA): MDText.com, Inc.; 2000-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK278977/

[4] R. C. Cervantes and U. M. Palacio, “Estimation of obesity levels based on Computational Intelligence,” *Informatics in Medicine Unlocked*, vol. 21, p. 100472, Oct. 2020.

[5] E. De-La-Hoz-Correa, F. E. Mendoza-Palechor, A. De-La-Hoz-Manotas, R. C. Morales-Ortega, and S. H. Beatriz Adriana, “Obesity level estimation software based on decision trees,” *Journal of Computer Science*, vol. 15, no. 1, pp. 67–77, Jan. 2019.

[6] C. Davila-Payan, M. DeGuzman, K. Johnson, N. Serban, and J. Swann, “Estimating prevalence of overweight or obese children and adolescents in small geographic areas using publicly available data,” *Centers for Disease Control and Prevention*, 15-Jun-2015. [Online]. Available: https://www.cdc.gov/pcd/issues/2015/14_0229.htm. [Accessed: 07-Feb-2023].

[7] X. Pang, C. B. Forrest, F. Lê-Scherban and A. J. Masino, "Understanding Early Childhood Obesity via Interpretation of Machine Learning Model Predictions," *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Boca Raton, FL, USA, 2019, pp. 1438-1443, doi: 10.1109/ICMLA.2019.00235.

[8] A. Choudhuri, “A hybrid machine learning model for estimation of obesity levels,” *Data Management, Analytics and Innovation*, vol. 137, pp. 315–329, Oct. 2022.

[9] R. Turner, D. Eriksson, M. McCourt, J. Kiili, E. Laaksonen, Z. Xu, and I. Guyon, “Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020,” *Proceedings of Machine Learning Research*, 20-Aug-2021. [Online]. Available: http://proceedings.mlr.press/v133/turner21a/turner21a.pdf. [Accessed: 07-Feb-2023].

[10] [D. Kirk, E. Kok, M. Tufano, B. Tekinerdogan, E. J. Feskens, and G. Camps, “Machine Learning in Nutrition Research,” *Advances in Nutrition*, vol. 13, no. 6, pp. 2573–2589, Sep. 2022.

[11] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions: Proceedings of the 31st International Conference on Neural Information Processing Systems,” *ACM Digital Library*, 01-Dec-2017. [Online]. Available: https://dl.acm.org/doi/10.5555/3295222.3295230. [Accessed: 07-Feb-2023].

[12] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, “From local explanations to global understanding with explainable AI for trees,” *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, Jan. 2020.