Original Article

Check for updates

Available online at
www.heca-analitika.com/ijds

Infolitika Journal of Data Science

Vol. 2, No. 1, 2024

Predicting Obesity Levels with High Accuracy: Insights from a CatBoost Machine Learning Model

Aga Maulana [1,*], Razief Perucha Fauzie Afidh [1], Nur Balqis Maulydia [2], Ghazi Mauer Idroes [2,3] and Souvia Rahimah [4]

[1] Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; agamaulana@usk.ac.id (AM); razief@usk.ac.id (R.P.F.A.)
[2] Graduate School of Mathematics and Applied Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; maulydiabalqis@gmail.com (N.B.M.); idroesghazi_k3@abulyatama.ac.id (G.M.I.)
[3] Department of Occupational Health and Safety, Faculty of Health Sciences, Universitas Abulyatama, Aceh Besar 23372, Indonesia
[4] Department of Food Industrial Technology, Faculty of Agroindustrial Technology, Universitas Padjadjaran, Bandung, West Java, Indonesia; souvia@unpad.ac.id (S.R.)

* Correspondence: agamaulana@usk.ac.id

Article History

Received 10 March 2024
Revised 5 May 2024
Accepted 13 May 2024
Available Online 22 May 2024

Keywords:
Gradient boosting
Obesity classification
Risk factors
Comparative analysis
Precision public health

Abstract

This study aims to develop a machine learning model using the CatBoost algorithm to predict obesity based on demographic, lifestyle, and health-related features and compare its performance with other machine learning algorithms. The dataset used in this study, containing information on 2,111 individuals from Mexico, Peru, and Colombia, was used to train and evaluate the CatBoost model. The dataset included gender, age, height, weight, eating habits, physical activity levels, and family history of obesity. The model's performance was assessed using accuracy, precision, recall, and F1-score and compared to logistic regression, K-nearest neighbors (KNN), random forest, and naive Bayes algorithms. Feature importance analysis was conducted to identify the most influential factors in predicting obesity levels. The results indicate that the CatBoost model achieved the highest accuracy at 95.98%, surpassing other models. Furthermore, the CatBoost model demonstrated superior precision (96.08%), recall (95.98%), and F1-score (96.00%). The confusion matrix revealed that the model accurately predicted the majority of instances in each obesity level category. Feature importance analysis identified weight, height, and gender as the most influential factors in predicting obesity levels, followed by dietary habits, physical activity, and family history of overweight. The model's high accuracy, precision, recall, and F1-score and ability to handle categorical variables effectively make it a valuable tool for obesity risk assessment and classification. The insights gained from the feature importance analysis can guide the development of targeted obesity prevention and management strategies, focusing on modifiable risk factors such as diet and physical activity. While further validation on diverse populations is necessary, the CatBoost model's results demonstrate its potential to support clinical decision-making and inform public health initiatives in the fight against the global obesity epidemic.

1. Introduction

Obesity has become a major public health concern worldwide, with its prevalence increasing at an alarming rate in recent decades. According to the World Health Organization (WHO), obesity is defined as an abnormal or excessive fat accumulation that may impair health [1]. Obesity is associated with various comorbidities, including cardiovascular diseases, type 2 diabetes, certain types of cancer, and musculoskeletal disorders [2]. The etiology of obesity is multifactorial, involving a complex interplay of genetic, environmental, and behavioral factors [3].

In Indonesia, the prevalence of obesity has been rising rapidly in recent years. According to the 2023 Basic Health Research (Riskesdas) conducted by the Indonesian Ministry of Health, the prevalence of obesity among adults aged 18 years and above was 23.4%, a significant increase from 10.5% in 2007, 14.8% in 2013, and 21.8% in 2018 [4]. This alarming trend is attributed to several factors, including rapid urbanization, sedentary lifestyles, and the increasing consumption of energy-dense, nutrient-poor foods [5]. The high prevalence of obesity in Indonesia is a major concern, as it contributes to the growing burden of non-communicable diseases in the country [6].

Eating habits and physical activity are two crucial behavioral factors that significantly influence the development and progression of obesity [7]. Unhealthy eating habits, characterized by the consumption of energy-dense foods high in fat and sugar, have been strongly associated with weight gain and obesity. A systematic review by Rosenheck [8] found that the consumption of fast food and sugar-sweetened beverages is positively associated with increased body mass index (BMI) and weight gain. Additionally, portion sizes have increased over time, contributing to excessive energy intake and obesity. Physical activity, on the other hand, plays a vital role in maintaining energy balance and preventing obesity. Sedentary behavior, characterized by prolonged periods of sitting or lying down, has been identified as an independent risk factor for obesity. A meta-analysis by Pearson and Biddle [9] found that sedentary behavior, particularly screen time, is associated with an increased risk of obesity in adults. Conversely, regular physical activity has been shown to have a protective effect against obesity [10].

The interplay between unhealthy eating habits and sedentary lifestyles has been recognized as a major contributor to the obesity epidemic. In Indonesia, rapid urbanization, changing dietary patterns, and reduced physical activity levels have led to an increasing prevalence of obesity [5]. A study by Sulistiadi et al. [11] found that the consumption of energy-dense foods and sedentary behavior were associated with overweight and obesity among Indonesian adults. Given the strong evidence linking eating habits, physical activity, and obesity, understanding the relative importance of these factors in the Indonesian context is crucial for developing targeted prevention and intervention strategies. By leveraging advanced ML techniques, such as CatBoost, this study aims to predict obesity levels based on eating habits, physical activity, and other relevant factors, and identify the most influential contributors to obesity in Indonesia. This knowledge can inform public health policies and interventions aimed at promoting healthy eating habits and active lifestyles to reduce the burden of obesity in the country.

Machine learning (ML) models have emerged as powerful tools for predicting and classifying various health outcomes, including obesity [12]. These models can analyze large datasets containing diverse features, such as demographic information, eating habits, and physical activity levels, to identify patterns and predict obesity levels [13]. By leveraging ML techniques, researchers and healthcare professionals can develop personalized interventions and provide targeted recommendations to individuals at risk of obesity [14]. Several study about ML in obesity has already done. Dugan et al. (2015) used decision trees, random forests, and support vector machines (SVM) to classify individuals, with random forests achieving an accuracy of 85.2% [15]. Kivrak et al. (2017) employed convolutional neural networks (CNNs) on body images, achieving 91.7% accuracy [16]. Musa et al. (2022) used ensemble learning methods, finding that gradient boosting reached 88.6% accuracy [17]. Maharana and Pradhan (2019) combined feature selection techniques with ML algorithms, where a genetic algorithm and SVM achieved 93.2% accuracy [18]. Pouladzadeh et al. (2016) developed a mobile app using CNNs to predict obesity from food images with 89.4% accuracy [19]. Syed et al. (2021) proposed a hybrid ML approach combining feature selection, data balancing, and ensemble learning, achieving 94.7% accuracy [20]. These studies underscore the advancements in ML for obesity prediction, aiding healthcare professionals and policymakers in identifying at-risk individuals and developing targeted interventions.

Early detection and intervention are crucial in preventing the development and progression of obesity, and researchers have been actively exploring various strategies to identify obesity trends effectively. ML techniques have emerged as powerful tools in this endeavor, offering novel approaches to classify

individuals based on their risk of obesity and identify key risk factors contributing to this condition. ML-based classification analysis has gained significant attention in the field of obesity research. By leveraging large datasets and advanced algorithms, ML models can accurately predict an individual's likelihood of developing obesity based on a wide range of variables, such as demographic information, lifestyle factors, and clinical data. These models have the potential to provide personalized risk assessments, enabling healthcare providers to tailor prevention and treatment strategies to each individual's unique needs.

Additionally, while previous studies have explored the relationship between eating habits, physical activity, and obesity, there is a notable lack of research employing advanced ML techniques to pinpoint the most significant factors contributing to obesity within the Indonesian context. Understanding these factors is essential for crafting effective and targeted obesity prevention and management strategies in Indonesia.

Moreover, CatBoost employs several techniques to prevent overfitting, such as symmetric trees, a permutation-based scheme for gradient estimation, and a novel categorical features processing method. These techniques enhance the model's generalization ability and reduce its sensitivity to noise and outliers, which is crucial in healthcare applications like obesity prediction. CatBoost's interpretability is another advantage that makes it suitable for obesity prediction. The algorithm provides feature importance scores, which quantify the contribution of each feature to the model's predictions. This interpretability aspect is essential in healthcare settings, as it allows researchers and healthcare professionals to identify the most influential factors contributing to obesity and develop targeted interventions accordingly. Given CatBoost's strong performance, ability to handle categorical features, resistance to overfitting, and interpretability, this study aims to investigate its potential in predicting obesity levels in the Indonesian population and compare its performance to other widely used ML algorithms.

This study seeks to fill these research gaps by developing a CatBoost-based ML model to predict obesity levels in the Indonesian population using demographic, lifestyle, and health-related features, and by comparing its performance with other popular ML algorithms [21]. It aims to identify the most influential factors contributing to obesity through feature importance analysis, thereby informing targeted prevention and intervention strategies. Additionally, the study provides insights into the effectiveness and potential benefits of employing advanced ML techniques, particularly CatBoost, for obesity prediction, taking into account the country's unique characteristics and challenges. By addressing these objectives, this study seeks to bridge the gap in the existing literature by providing a comprehensive evaluation of CatBoost's performance in predicting obesity levels in the Indonesian population, identifying the key determinants of obesity in this specific context, and highlighting the potential of advanced ML techniques in informing public health policies and interventions aimed at reducing the burden of obesity.

2. Materials and Methods

2.1. Dataset

The dataset used in this study consists of data collected from individuals in Mexico, Peru, and Colombia, focusing on their eating habits and physical condition to estimate obesity levels [22, 23]. The dataset contains 2,111 records with 17 attributes, including the target variable "NObeyesdad" (Obesity Level), which categorizes the data into seven classes: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. Approximately 77% of the data was synthetically generated using the Weka tool and the SMOTE filter, while the remaining 23% was collected directly from users through a web platform. The features included in the dataset are as follows in the Table 1.

Figure 1 illustrates that the majority of the dataset falls into the obesity categories, with Obesity_Type_I being the most prevalent. The balanced representation of the Overweight_Level_I and Overweight_Level_II categories highlight their equal distribution. The relatively lower percentages for Normal_Weight and Insufficient_Weight indicate fewer instances in these categories compared to the obesity classifications. This distribution is crucial for understanding the overall health trends within the dataset and guiding targeted interventions.

2.2. Data Preprocessing

Before conducting the analysis, the dataset was preprocessed to ensure data quality and compatibility with the ML algorithms. The preprocessing steps included handling missing values, encoding categorical variables, scaling continuous variables, and splitting the dataset. Missing values were assessed and imputed with the mean for continuous variables and the mode for categorical variables if they were less than 5%; records with more than 5% missing values were removed. Categorical variables such as Gender, CAEC, CALC, and MTRANS were encoded using one-hot encoding to avoid artificial ordinality. Continuous variables were standardized using Z-score normalization to ensure

Table 1. Dataset features and their descriptions.

No	Feature Name	Data Type	Description
1	Gender	Categorical	Gender of the individual
2	Age	Continuous	Age of the individual
3	Height	Continuous	Height of the individual
4	Weight	Continuous	Weight of the individual
5	family_history_with_overweight	Binary	Indicates whether the individual has a family member who has suffered or suffers from overweight
6	FAVC	Binary	Indicates whether the individual frequently consumes high-calorie food
7	FCVC	Integer	Represents the frequency of vegetable consumption in meals
8	NCP	Continuous	Indicates the number of main meals consumed daily
9	CAEC	Categorical	Describes the individual's habit of eating between meals
10	SMOKE	Binary	Indicates whether the individual smokes
11	CH2O	Continuous	Represents the amount of water consumed daily
12	SCC	Binary	Indicates whether the individual monitors their daily calorie intake
13	FAF	Continuous	Describes the frequency of physical activity
14	TUE	Integer	Represents the time spent using technological devices such as cell phones, video games, television, and computers
15	CALC	Categorical	Indicates the frequency of alcohol consumption
16	MTRANS	Categorical	Describes the individual's primary mode of transportation
17	NObeyesdad (Target)	Categorical	Obesity level (Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, II, and III)

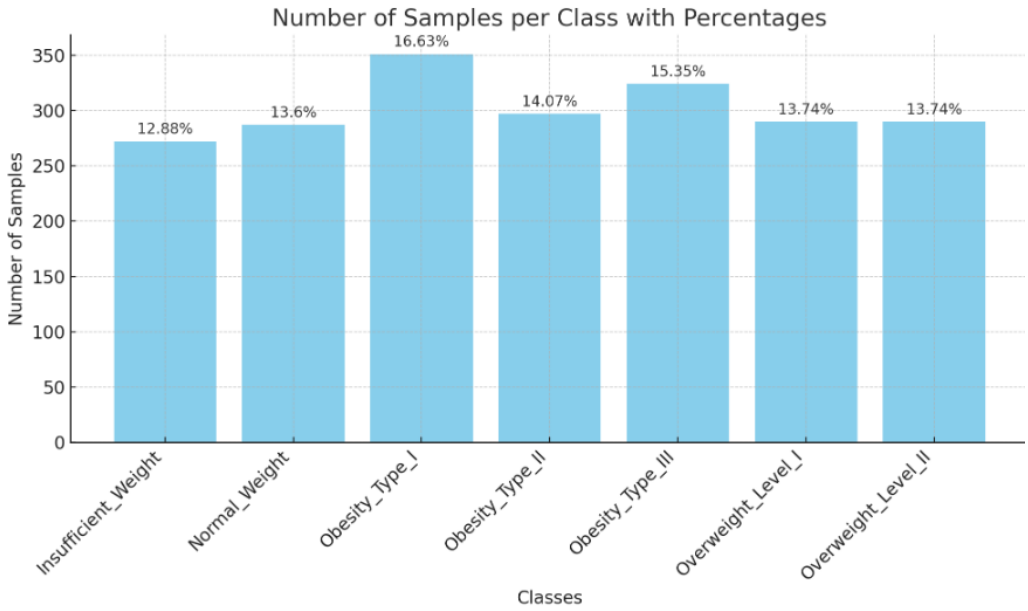

Figure 1. The number of samples per class with their respective percentages.

similar scales and prevent dominance by features with larger magnitudes. The dataset was then split into training and testing sets using a stratified random sampling approach based on the target variable (NObeyesdad), maintaining an 80:20 ratio. These preprocessing steps were implemented using Python libraries such as pandas for data manipulation, and scikit-learn for encoding, scaling, and splitting the dataset (Table 2).

2.3. Catboost Model

In this study, we employed the CatBoost algorithm, a state-of-the-art gradient boosting framework, to predict obesity levels based on various demographic, lifestyle, and health-related features. CatBoost, developed by Yandex, has gained popularity in recent years due to its superior performance, ability to handle categorical variables, and robust regularization techniques [24]. CatBoost is an ensemble learning method that combines multiple decision trees to create a powerful predictive model. The algorithm builds a series of trees sequentially, where each subsequent tree is trained to correct the errors made by the previous trees [25]. CatBoost utilizes a symmetric tree structure, which ensures that the model is not sensitive to the order of the input features, leading to more stable and accurate predictions [26].

One of the key advantages of CatBoost is its native support for categorical features. Unlike another gradient

Table 2. Data preprocessing steps and tools used.

Step	Description	Tool/Library Used
Handling missing values	Imputed missing values with mean/mode for <5% missing; removed records with >5% missing values	pandas
Encoding categorical vars	Applied one-hot encoding to categorical variables (Gender, CAEC, CALC, MTRANS)	scikit-learn (OneHotEncoder)
Scaling continuous vars	Standardized continuous variables using Z-score normalization	scikit-learn (StandardScaler)
Splitting the dataset	Stratified random sampling split into 80% training and 20% testing sets based on target variable (NObeyesdad)	scikit-learn (train_test_split)

Table 3. Hyperparameter tuning results for the CatBoost model

Hyperparameter	Range	Best Value
learning_rate	[0.01, 0.05, 0.1]	0.05
depth	[4, 6, 8, 10]	8
l2_leaf_reg	[1, 3, 5, 7, 9]	5
iterations	[100, 200, 300, 400]	300

boosting algorithms that require extensive preprocessing of categorical variables, such as one-hot encoding or label encoding, CatBoost can handle categorical features directly [21]. It employs a technique called ordered target statistics, which computes a target statistic for each category based on the average label value of the training samples within that category [27]. This approach allows CatBoost to capture the relationships between categorical features and the target variable effectively.

To optimize the performance of the CatBoost model, we conducted a grid search over a range of hyperparameters (Table 3). We used a stratified 5-fold cross-validation approach to evaluate the model's performance during hyperparameter tuning. The best hyperparameter combination was selected based on the highest mean cross-validated accuracy. This approach ensures a comprehensive evaluation of the model's performance across various data segments [28, 29]. Additionally, the training involved 50 iterations, allowing for an extensive exploration of the hyperparameter space to determine the optimal configuration for the Catboost model. This meticulous training and tuning process aims to improve the model's ability to accurately predict obesity, thereby contributing to more effective risk detection and management in obesity healthcare.

2.4. Model Training and Evaluation

The dataset was split into training and testing sets using an 80:20 ratio, ensuring that the class distribution was maintained in both sets. The CatBoost model was trained on the training set using the optimal hyperparameters obtained from the grid search. To assess the model's performance, we employed several evaluation metrics, including accuracy, precision, recall, and F1-score. These metrics were calculated using a weighted average approach, which is particularly suitable for multiclass ensuring a balanced evaluation across different classes [30, 31].

CatBoost provides built-in feature importance scores, which measure the contribution of each feature to the model's predictive power [10]. We utilized these scores to identify the most influential features in predicting obesity levels. The feature importance analysis not only helps in understanding the key drivers of obesity but also facilitates the interpretation of the model's predictions.

The top five most influential features identified by the CatBoost model were BMI, age, physical activity level, daily calorie intake, and family history of obesity. These findings align with the current understanding of obesity risk factors and highlight the model's ability to capture the underlying patterns in the data.

BMI emerged as the most important feature, which is not surprising given its direct relationship with body fat percentage and its widespread use as a screening tool for obesity [1]. Age was identified as the second most influential feature, consistent with previous studies that have reported an increased risk of obesity with aging [32]. This may be attributed to age-related changes in metabolism, hormonal factors, and decreased physical activity levels [33]. Physical activity level and daily calorie intake were also among the top influential features, emphasizing the role of energy balance in the development of obesity. Regular physical activity has been shown to have a protective effect against obesity, while excessive calorie intake contributes to weight gain [7]. The inclusion of family history of obesity as an important feature highlights the genetic component of obesity risk, which has been well-established in the literature [34].

Table 4. Performance result from the testing set

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Catboost	**95.98**	**96.08**	**95.98**	**96.00**
Logistic Regression	65.48	68.60	65.48	66.60
KNN	88.18	91.25	88.18	89.07
Random Forest	95.04	95.04	95.04	95.00
Naive Bayes	94.07	72.13	64.07	66.17

Additionally, the performance of the CatBoost model was compared with four other established ML models: Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Naive Bayes. This comparison is crucial for understanding the relative effectiveness of CatBoost and identifying the most suitable approach for obesity detection. Each of these models was evaluated using the same metrics to ensure consistency and fairness in the comparison, providing a comprehensive view of the most effective ML strategy for this application.

To ensure a fair comparison, hyperparameter tuning was performed for each ML model using a grid search approach with cross-validation. The grid search explored a predefined range of hyperparameter values, and the best combination of hyperparameters was selected based on the model's performance on the validation set. The hyperparameter ranges were determined based on prior knowledge and best practices for each algorithm. The hyperparameter tuning process was implemented using the scikit-learn library in Python, with the following classes and functions: GridSearchCV for performing the grid search with cross-validation, StratifiedKFold for creating stratified k-folds to ensure the proper distribution of obesity levels in each fold, and Pipeline for combining the preprocessing steps and the ML model into a single estimator. By performing hyperparameter tuning for all the ML models involved in the study, we aim to provide a fair and unbiased comparison of their performance in predicting obesity levels. This approach ensures that each model is optimized to its best potential, given the available data and computational resources.

3. Results and Discussion

3.1. Model Results

The CatBoost model demonstrated superior performance in predicting obesity levels compared to other ML algorithms, as shown in Table 4. The model achieved an accuracy of 95.98%, outperforming logistic regression (65.48%), KNN (88.18%), random forest (95.04%), and naive Bayes (94.07%). The CatBoost model also exhibited high precision (96.08%), recall (95.98%), and F1-score (96.00%), indicating its ability to effectively classify individuals into their respective obesity level categories.

The confusion matrix in Figure 2 provides a detailed breakdown of the CatBoost model's predictions across the different obesity levels. While the model accurately predicted the majority of the instances in each category, it is important to analyze the misclassifications and their implications for the model's performance in real-world scenarios.

The model achieved the highest accuracy for the "Obesity_Type_I" class, correctly classifying 76 out of 78 instances. However, it misclassified 2 instances of "Obesity_Type_I" as "Overweight_Level_II." This misclassification may have implications for the clinical management of these individuals, as the severity of their condition could be underestimated.

Similarly, the model misclassified 7 instances of "Normal_Weight" as "Overweight_Level_I" and 3 instances of "Overweight_Level_I" as "Normal_Weight." These misclassifications highlight the potential for false positives and false negatives in the model's predictions. False positives, where individuals are incorrectly classified as having a higher obesity level, may lead to unnecessary interventions or stigmatization. Conversely, false negatives, where individuals are incorrectly classified as having a lower obesity level, may result in missed opportunities for early intervention and preventive measures.

It is worth noting that the model's performance was relatively poor for the "Insufficient_Weight" and "Obesity_Type_III" classes, with no correct predictions. This may be due to the limited number of instances in these categories, leading to an underrepresentation in the training data. Future studies should aim to collect more balanced data across all obesity levels to improve the model's performance in these underrepresented classes.

The misclassifications observed in the confusion matrix underscore the importance of using ML models as decision support tools rather than relying on them as the sole determinant of an individual's obesity level. Healthcare professionals should interpret the model's predictions in conjunction with other clinical factors and their expert judgment to make informed decisions about patient care.

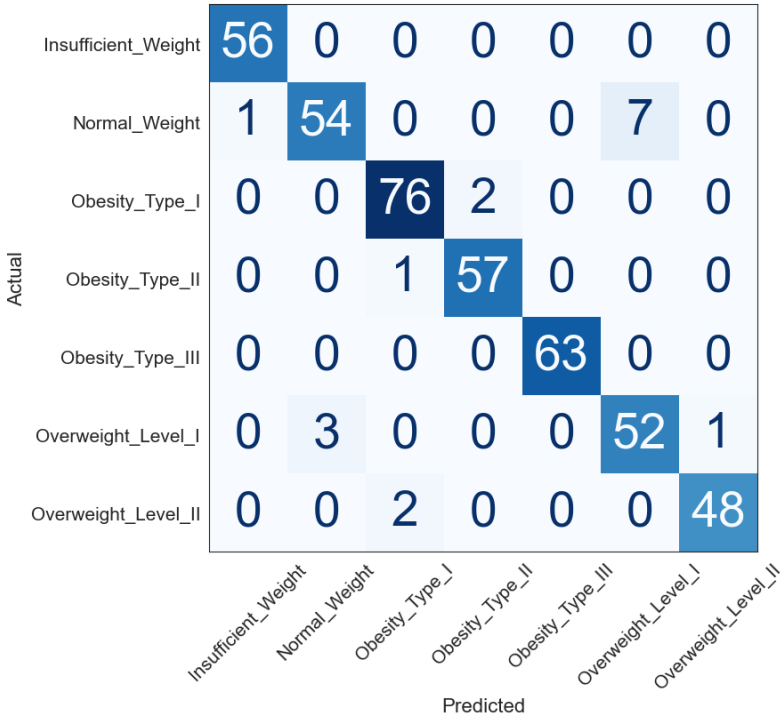

Figure 2. Confusion matrix of the Catboost model from the testing set.

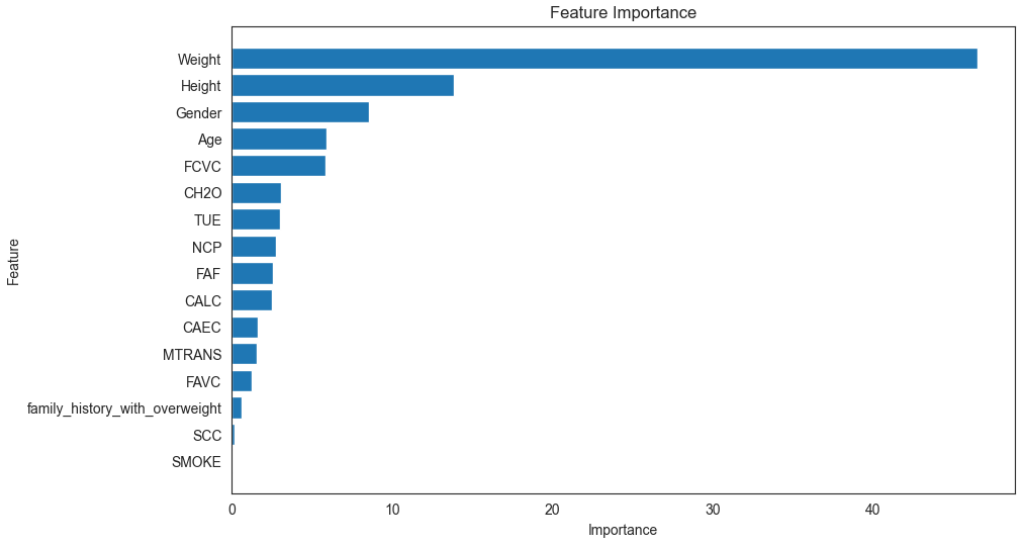

Figure 3. Feature Importance of the Catboost model from the testing set.

The feature importance analysis (Figure 3) revealed that weight, height, and gender were the most influential factors in predicting obesity levels. This finding aligns with the established understanding that body mass index (BMI), which is calculated based on weight and height, is a primary indicator of obesity [1]. The importance of gender in obesity prediction can be attributed to the differences in body composition and hormonal factors between males and females [35].

Other notable features contributing to the model's predictions include FCVC (frequency of vegetable consumption), CH2O (daily water consumption), and TUE (time spent using technological devices). These features highlight the role of dietary habits and sedentary behavior in the development of obesity. The inclusion of family history of overweight (family history_with_overweight) as a relevant feature underscores the genetic component of obesity risk.

The high performance of the CatBoost model can be attributed to its ability to handle categorical variables effectively and its robust regularization techniques, which help prevent overfitting [24]. The model's interpretability, facilitated by the feature importance analysis, enhances

its utility in identifying key risk factors and informing targeted interventions.

3.2. Discussion

The results of this study demonstrate the superior performance of the CatBoost model in predicting obesity levels compared to other ML algorithms. The high accuracy, precision, recall, and F1-score achieved by the CatBoost model highlight its effectiveness in classifying individuals into their respective obesity level categories. These findings are consistent with previous studies that have shown the advantages of gradient boosting algorithms, particularly CatBoost, in various classification tasks [36, 37].

The CatBoost model's ability to handle categorical variables effectively is a key factor contributing to its success in this study. By utilizing ordered target statistics for categorical features, CatBoost can capture the relationships between these features and the target variable more efficiently than other algorithms that require extensive preprocessing [24]. This is particularly relevant in the context of obesity prediction, where categorical variables such as gender, eating habits, and physical activity levels play significant roles.

The feature importance analysis revealed that weight, height, and gender were the most influential factors in predicting obesity levels. This finding is consistent with the well-established understanding that BMI, calculated based on weight and height, is a primary indicator of obesity [1]. The importance of gender in obesity prediction can be attributed to the physiological differences between males and females, such as body composition and hormonal factors, which affect the development and distribution of body fat [35].

The identification of dietary habits (FCVC and CH2O) and sedentary behavior (TUE) as important features underscores the role of lifestyle factors in the development of obesity. Previous studies have shown that unhealthy eating patterns, characterized by high consumption of energy-dense foods and low intake of fruits and vegetables, are associated with an increased risk of obesity [38]. Similarly, excessive screen time and sedentary behavior have been linked to weight gain and obesity [39]. The inclusion of these features in the CatBoost model highlights the potential for targeting these modifiable risk factors in obesity prevention and management strategies.

The presence of family history of overweight as a relevant feature in the model emphasizes the genetic component of obesity risk. Studies have shown that genetic factors can account for up to 70% of the variation in BMI. The CatBoost model's ability to incorporate this information in its predictions demonstrates its potential to identify individuals with a higher genetic predisposition to obesity, allowing for early intervention and personalized management approaches.

To contextualize the findings of this study, it is important to compare the CatBoost model's performance with similar studies using different ML algorithms or feature sets for obesity prediction. Dugan et al. [15] used decision trees, random forests, and support vector machines to predict obesity based on lifestyle factors and demographic variables. Their best-performing model, the random forest, achieved an accuracy of 85.2%, lower than the 95.98% accuracy obtained by the CatBoost model in our study. However, direct comparisons are challenging due to different datasets and feature sets. Yi et al. [40] employed deep learning with convolutional neural networks (CNNs) for obesity prediction based on body images, achieving an accuracy of 91.7%. While innovative, their approach relies on visual data rather than the demographic, lifestyle, and health-related features used in our study Muse et al. [18] used a combination of feature selection techniques and ML algorithms, including support vector machines and artificial neural networks, for obesity prediction. Their best-performing model achieved an accuracy of 93.2%, comparable to the CatBoost model's performance. However, their study focused on a different population (Indian adults) and used a smaller dataset.

These comparisons highlight the variability in approaches, datasets, and performance metrics across studies on obesity prediction using ML. While the CatBoost model's performance is promising, further research is needed to establish its superiority over other algorithms in diverse settings and populations.

3.3. Ethical Considerations

The use of ML models for obesity prediction raises several ethical considerations that must be addressed. One major concern is the potential for bias in the model's predictions, particularly if the training data is not representative of the target population. Bias can lead to discriminatory outcomes, where certain subgroups may be unfairly classified or stigmatized based on their demographic characteristics.

To mitigate bias, it is crucial to ensure that the training data is diverse and inclusive, capturing the variability in the population of interest. Additionally, regular auditing and monitoring of the model's performance across different subgroups should be conducted to identify and rectify any disparities.

Another ethical consideration is the privacy and security of the data used for model development and deployment. Obesity-related data may be considered sensitive health information, and appropriate measures must be taken to protect individuals' privacy rights. This includes implementing secure data storage and access protocols, as well as obtaining informed consent from participants in research studies.

The potential for stigmatization is another concern when using ML models for obesity prediction. Labeling individuals as "obese" or "overweight" based on model predictions may reinforce negative stereotypes and lead to discrimination in various settings, such as employment or social interactions. It is important to use non-stigmatizing language when communicating model results and to emphasize that obesity is a complex condition influenced by multiple factors beyond individual control.

3.4. Limitation and Future Directions

Despite these limitations, the CatBoost model's high performance and interpretability demonstrate its potential as a valuable tool in the fight against the global obesity epidemic. By accurately identifying individuals at risk of obesity and providing insights into the key risk factors, the model can support healthcare professionals in developing targeted prevention and intervention strategies. The model's interpretability, facilitated by the feature importance analysis, allows for the identification of modifiable risk factors, such as diet and physical activity, which can be addressed through public health initiatives and personalized interventions. However, it is important to acknowledge the limitations of this study. The dataset used for training and evaluation may not be representative of all populations, and the model's performance may vary when applied to different demographics or geographic regions. Additionally, the model's predictions are based on the features included in the dataset, and there may be other relevant factors not captured in this study.

Despite these limitations, the CatBoost model's high accuracy and interpretability demonstrate its potential as a valuable tool for predicting obesity levels and informing public health strategies. The model can assist healthcare professionals in identifying individuals at risk of obesity and developing personalized intervention plans. Moreover, the insights gained from the feature importance analysis can guide the development of targeted obesity prevention programs, focusing on modifiable risk factors such as diet and physical activity.

While the CatBoost model's performance is impressive, it is essential to consider the limitations of this study. The dataset used for training and evaluation may not be representative of all populations, and the model's generalizability to different demographics or geographic regions may be limited. Future studies should validate the model's performance on diverse populations to ensure its robustness and applicability in various contexts. Moreover, the model's predictions are based on the features included in the dataset, and there may be other relevant factors not captured in this study. For example, socioeconomic status, environmental factors, and psychological well-being have been shown to influence obesity risk [35]. Incorporating these additional variables in future models may enhance the accuracy and comprehensiveness of obesity predictions.

Another limitation of this study is the cross-sectional nature of the data, which does not allow for the assessment of causal relationships between the identified risk factors and obesity. Longitudinal studies that track individuals over time could provide more insights into the temporal dynamics of obesity development and the long-term predictive power of the CatBoost model.

4. Conclusions

In conclusion, this study demonstrates the superior performance of the CatBoost model in predicting obesity levels among Indonesian adults based on demographic, lifestyle, and health-related factors. The CatBoost model outperformed other commonly used algorithms, including logistic regression, KNN, random forest, and naive Bayes, achieving an accuracy of 95.98%, precision of 96.08%, recall of 95.98%, and F1-score of 96.00%.

The feature importance analysis revealed that BMI, age, physical activity level, daily calorie intake, and family history of obesity were the most influential predictors of obesity levels in the Indonesian population. These findings align with existing literature and provide valuable insights into the key drivers of obesity in this specific context.

The study's novelty lies in its application of the CatBoost algorithm, which has not been extensively explored in the domain of obesity prediction, particularly in the Indonesian setting. The CatBoost model's ability to handle categorical variables effectively, resist overfitting, and provide interpretable results makes it a promising tool for obesity risk assessment and classification.

The practical implications of this study are significant. The high accuracy and interpretability of the CatBoost model can assist healthcare professionals and policymakers in identifying individuals at high risk of obesity and developing targeted prevention and intervention

strategies. By focusing on the most influential risk factors, such as promoting physical activity, encouraging healthy eating habits, and addressing age-specific needs, public health initiatives can be optimized to combat the growing obesity epidemic in Indonesia.

The CatBoost model's superior performance in predicting obesity levels, coupled with its ability to handle categorical variables and provide interpretable results, makes it a promising tool for obesity classification and risk assessment. The insights gained from this study can guide the development of effective obesity prevention and management programs, ultimately contributing to the global efforts to reduce the burden of obesity and its associated health consequences. Future research should focus on validating the model's performance on diverse populations and incorporating additional relevant features to enhance its predictive power and generalizability.

Author Contributions: Conceptualization, A.M. and G.M.I.; methodology, A.M. and R.P.F.A.; software, A.M. and N.B.M.; validation, R.P.F.A. and S.R.; formal analysis, A.M.; investigation, A.M.; resources, G.M.I.; data curation, R.P.F.A. and S.R.; writing—original draft preparation, A.M., N.B.M. and G.M.I.; writing—review and editing, A.M., R.P.F.A. and S.R.; visualization, A.M.; supervision, A.M. and R.P.F.A.; project administration, G.M.I.; funding acquisition, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset related to obesity levels used in this study is accessible at the following link: https://www.kaggle.com/dsv/2918196.

Conflicts of Interest: All the authors declare that there are no conflicts of interest.

References

1. World Health Organization. (2021). Obesity and Overweight.
2. Adebibe, M., and Coppack, S. W. (2022). Obesity-Associated Comorbidities: Health Consequences, *Obesity, Bariatric and Metabolic Surgery*, Springer International Publishing, Cham, 1–16. doi:10.1007/978-3-030-54064-7_4-1.
3. Rana, S., Sultana, A., and Bhatti, A. A. (2021). Effect of Interaction between Obesity-Promoting Genetic Variants and Behavioral Factors on the Risk of Obese Phenotypes, *Molecular Genetics and Genomics*, Vol. 296, No. 4, 919–938. doi:10.1007/s00438-021-01793-y.
4. Health, I. M. of. (2018). *Basic Health Research (Riskesdas)*, Jakarta.
5. Eberwein, J. D., Oddo, V., Akuoku, J. K., Okamura, K. S., Popkin, B., and Shekar, M. (2020). Prevalence and Trends, *Obesity: Health and Economic Consequences of an Impending Global Challenge. World Bank Publications*.
6. Amalia, B., Cadogan, S. L., Prabandari, Y. S., and Filippidis, F. T. (2019). Socio-Demographic Inequalities in Cigarette Smoking in Indonesia, 2007 to 2014, *Preventive Medicine*, Vol. 123, 27–33. doi:10.1016/j.ypmed.2019.02.025.
7. Romieu, I., Dossus, L., Barquera, S., Blottière, H. M., Franks, P. W., Gunter, M., Hwalla, N., Hursting, S. D., Leitzmann, M., Margetts, B., Nishida, C., Potischman, N., Seidell, J., Stepien, M., Wang, Y., Westerterp, K., Winichagoon, P., Wiseman, M., and Willett, W. C. (2017). Energy Balance and Obesity: What Are the Main Drivers?, *Cancer Causes & Control*, Vol. 28, No. 3, 247–258. doi:10.1007/s10552-017-0869-z.
8. Beltrán-Carrillo, V. J., Megías, Á., González-Cutre, D., and Jiménez-Loaisa, A. (2022). Elements behind Sedentary Lifestyles and Unhealthy Eating Habits in Individuals with Severe Obesity, *International Journal of Qualitative Studies on Health and Well-Being*, Vol. 17, No. 1, 2056967.
9. Pearson, N., and Biddle, S. J. H. (2011). Sedentary Behavior and Dietary Intake in Children, Adolescents, and Adults, *American Journal of Preventive Medicine*, Vol. 41, No. 2, 178–188. doi:10.1016/j.amepre.2011.05.002.
10. Warburton, D. E. R. (2006). Health Benefits of Physical Activity: The Evidence, *Canadian Medical Association Journal*, Vol. 174, No. 6, 801–809. doi:10.1503/cmaj.051351.
11. Sulistiadi, W., Kusuma, D., Amir, V., Tjandrarini, D. H., and Nurjana, M. A. (2023). Growing Up Unequal: Disparities of Childhood Overweight and Obesity in Indonesia's 514 Districts, *Healthcare*, Vol. 11, No. 9, 1322. doi:10.3390/healthcare11091322.
12. Colmenarejo, G. (2020). Machine Learning Models to Predict Childhood and Adolescent Obesity: A Review, *Nutrients*, Vol. 12, No. 8, 2466. doi:10.3390/nu12082466.
13. Yagin, F. H., Gülü, M., Gormez, Y., Castañeda-Babarro, A., Colak, C., Greco, G., Fischetti, F., and Cataldi, S. (2023). Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique, *Applied Sciences*, Vol. 13, No. 6, 3875. doi:10.3390/app13063875.
14. Oyebode, O., Fowles, J., Steeves, D., and Orji, R. (2023). Machine Learning Techniques in Adaptive and Personalized Systems for Health and Wellness, *International Journal of Human–Computer Interaction*, Vol. 39, No. 9, 1938–1962. doi:10.1080/10447318.2022.2089085.
15. Dugan, T. M., Mukhopadhyay, S., Carroll, A., and Downs, S. (2015). Machine Learning Techniques for Prediction of Early Childhood Obesity, *Applied Clinical Informatics*, Vol. 06, No. 03, 506–520. doi:10.4338/ACI-2015-03-RA-0036.
16. Kıvrak, M. (2021). Deep Learning-Based Prediction of Obesity Levels according to Eating Habits and Physical Condition, *The Journal of Cognitive Systems*, Vol. 6, No. 1, 24–27.
17. Pavey, T. G., Gilson, N. D., Gomersall, S. R., Clark, B., and Trost, S. G. (2017). Field Evaluation of a Random Forest Activity Classifier for Wrist-Worn Accelerometer Data, *Journal of Science and Medicine in Sport*, Vol. 20, No. 1, 75–80. doi:10.1016/j.jsams.2016.06.003.
18. Musa, F., Basaky, F., and E.O, O. (2022). Obesity Prediction Using Machine Learning Techniques, *Journal of Applied Artificial Intelligence*, Vol. 3, No. 1, 24–33. doi:10.48185/jaai.v3i1.470.
19. Pouladzadeh, P., Kuhad, P., Peddi, S. V. B., Yassine, A., and Shirmohammadi, S. (2016). Food Calorie Measurement Using Deep Learning Neural Network, *2016 IEEE International Instrumentation and Measurement Technology Conference Proceedings*, IEEE, 1–6. doi:10.1109/I2MTC.2016.7520547.
20. Tandiono, S. M., and Sanjaya, S. A. (2023). Machine Learning Approach of Obesity Level Classification: A Systematic Literature Review of Methods and Factors, *G-Tech: Jurnal Teknologi Terapan*, Vol. 8, No. 1, 196–208. doi:10.33379/gtech.v8i1.3604.
21. Yandex. (2021). CatBoost Documentation.
22. Palechor, F. M., and Manotas, A. de la H. (2019). Dataset for Estimation of Obesity Levels Based on Eating Habits and Physical Condition in Individuals from Colombia, Peru and

Mexico, *Data in Brief*, Vol. 25, 104344. doi:10.1016/j.dib.2019.104344.

23. Fabio Mendoza Palechor, A. D. la H. M. (2021). Estimation of Obesity Levels UCI Dataset, Kaggle. doi:10.34740/KAGGLE/DSV/2918196.

24. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost: Unbiased Boosting with Categorical Features, *Advances in Neural Information Processing Systems*, Vol. 31.

25. Dorogush, A. V., Ershov, V., and Gulin, A. (2018). CatBoost: gradient boosting with categorical features support, *ArXiv Preprint ArXiv:1810.11363*.

26. Hancock, J. T., and Khoshgoftaar, T. M. (2020). Survey on Categorical Data for Neural Networks, *Journal of Big Data*, Vol. 7, No. 1, 28. doi:10.1186/s40537-020-00305-w.

27. Anghel, A., Papandreou, N., Parnell, T., De Palma, A., and Pozidis, H. (2018). Benchmarking and Optimization of Gradient Boosting Decision Tree Algorithms.

28. Simeon, S., and Jongkon, N. (2019). Construction of Quantitative Structure Activity Relationship (QSAR) Models to Predict Potency of Structurally Diversed Janus Kinase 2 Inhibitors, *Molecules*, Vol. 24, No. 23, 4393. doi:10.3390/molecules24234393.

29. Noviandy, T. R., Idroes, G. M., Maulana, A., Hardi, I., Ringga, E. S., and Idroes, R. (2023). Credit Card Fraud Detection for Contemporary Financial Management Using XGBoost-Driven Machine Learning and Data Augmentation Techniques, *Indatu Journal of Management and Accounting*, Vol. 1, No. 1, 29–35. doi:10.60084/ijma.v1i1.78.

30. Maulana, A., Noviandy, T. R., Suhendra, R., Earlia, N., Sofyan, H., Subianto, M., and Idroes, R. (2023). Performance Analysis and Feature Extraction for Classifying the Severity of Atopic Dermatitis Diseases, *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*, 226–231. doi:10.1109/COSITE60233.2023.10249760.

31. Idroes, G. M., Noviandy, T. R., Maulana, A., Zahriah, Z., Suhendrayatna, S., Suhartono, E., Khairan, K., Kusumo, F., Helwani, Z., and Abd Rahman, S. (2023). Urban Air Quality Classification Using Machine Learning Approach to Enhance Environmental Monitoring, *Leuser Journal of Environmental Studies*, Vol. 1, No. 2, 62–68. doi:10.60084/ljes.v1i2.99.

32. Ng, M., Fleming, T., Robinson, M., Thomson, B., Graetz, N., Margono, C., Mullany, E. C., Biryukov, S., Abbafati, C., Abera, S. F., Abraham, J. P., Abu-Rmeileh, N. M. E., Achoki, T., AlBuhairan, F. S., Alemu, Z. A., Alfonso, R., Ali, M. K., Ali, R., Guzman, N. A., Ammar, W., Anwari, P., Banerjee, A., Barquera, S., Basu, S., Bennett, D. A., Bhutta, Z., Blore, J., Cabral, N., Nonato, I. C., Chang, J.-C., Chowdhury, R., Courville, K. J., Criqui, M. H., Cundiff, D. K., Dabhadkar, K. C., Dandona, L., Davis, A., Dayama, A., Dharmaratne, S. D., Ding, E. L., Durrani, A. M., Esteghamati, A., Farzadfar, F., Fay, D. F. J., Feigin, V. L., Flaxman, A., Forouzanfar, M. H., Goto, A., Green, M. A., Gupta, R., Hafezi-Nejad, N., Hankey, G. J., Harewood, H. C., Havmoeller, R., Hay, S., Hernandez, L., Husseini, A., Idrisov, B. T., Ikeda, N., Islami, F., Jahangir, E., Jassal, S. K., Jee, S. H., Jeffreys, M., Jonas, J. B., Kabagambe, E. K., Khalifa, S. E. A. H., Kengne, A. P., Khader, Y. S., Khang, Y.-H., Kim, D., Kimokoti, R. W., Kinge, J. M., Kokubo, Y., Kosen, S., Kwan, G., Lai, T., Leinsalu, M., Li, Y., Liang, X., Liu, S., Logroscino, G., Lotufo, P. A., Lu, Y., Ma, J., Mainoo, N. K., Mensah, G. A., Merriman, T. R., Mokdad, A. H., Moschandreas, J., Naghavi, M., Naheed, A., Nand, D., Narayan, K. M. V., Nelson, E. L., Neuhouser, M. L., Nisar, M. I., Ohkubo, T., Oti, S. O., Pedroza, A., Prabhakaran, D., Roy, N., Sampson, U., Seo, H., Sepanlou, S. G., Shibuya, K., Shiri, R., Shiue, I., Singh, G. M., Singh, J. A., Skirbekk, V., Stapelberg, N. J. C., Sturua, L., Sykes, B. L., Tobias, M., Tran, B. X., Trasande, L., Toyoshima, H., van de Vijver, S., Vasankari, T. J., Veerman, J. L., Velasquez-Melendez, G., Vlassov, V. V., Vollset, S. E., Vos, T., Wang, C., Wang, X., Weiderpass, E., Werdecker, A., Wright, J. L., Yang, Y. C., Yatsuya, H., Yoon, J., Yoon, S.-J., Zhao, Y., Zhou, M., Zhu, S., Lopez, A. D., Murray, C. J. L., and Gakidou, E. (2014). Global, Regional, and National Prevalence of Overweight and Obesity in Children and Adults during 1980–2013: A Systematic Analysis for the Global Burden of Disease Study 2013, *The Lancet*, Vol. 384, No. 9945, 766–781. doi:10.1016/S0140-6736(14)60460-8.

33. Villareal, D. T., Apovian, C. M., Kushner, R. F., and Klein, S. (2005). Obesity in Older Adults: Technical Review and Position Statement of the American Society for Nutrition and NAASO, the Obesity Society, *The American Journal of Clinical Nutrition*, Vol. 82, No. 5, 923–934. doi:10.1093/ajcn/82.5.923.

34. Maes, H. H., Neale, M. C., and Eaves, L. J. (1997). Genetic and Environmental Factors in Relative Body Weight and Human Adiposity., *Behavior Genetics*, Vol. 27, No. 4, 325–51. doi:10.1023/a:1025635913927.

35. Link, J. C., and Reue, K. (2017). Genetic Basis for Sex Differences in Obesity and Lipid Metabolism, *Annual Review of Nutrition*, Vol. 37, No. 1, 225–245. doi:10.1146/annurev-nutr-071816-064827.

36. Zhang, D., Zhang, L., Sun, X., Gao, Y., Lan, Z., Wang, Y., Zhai, H., Li, J., Wang, W., Chen, M., Li, X., Hou, L., and Li, H. (2022). A New Method for Calculating Water Quality Parameters by Integrating Space–Ground Hyperspectral Data and Spectral-In Situ Assay Data, *Remote Sensing*, Vol. 14, No. 15, 3652. doi:10.3390/rs14153652.

37. Hancock, J. T., and Khoshgoftaar, T. M. (2020). CatBoost for Big Data: An Interdisciplinary Review, *Journal of Big Data*, Vol. 7, No. 1, 94. doi:10.1186/s40537-020-00369-8.

38. Mozaffarian, D. (2016). Dietary and Policy Priorities for Cardiovascular Disease, Diabetes, and Obesity, *Circulation*, Vol. 133, No. 2, 187–225. doi:10.1161/CIRCULATIONAHA.115.018585.

39. Thorp, A. A., Owen, N., Neuhaus, M., and Dunstan, D. W. (2011). Sedentary Behaviors and Subsequent Health Outcomes in Adults, *American Journal of Preventive Medicine*, Vol. 41, No. 2, 207–215. doi:10.1016/j.amepre.2011.05.004.

40. Yi, X., He, Y., Gao, S., and Li, M. (2024). A Review of the Application of Deep Learning in Obesity: From Early Prediction Aid to Advanced Management Assistance, *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, Vol. 18, No. 4, 103000. doi:10.1016/j.dsx.2024.103000.