ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Comparison of synthetic data generation techniques for obesity level prediction based on dietary habits and physical status

Hakan Alp Eren[1], Halil İbrahim Emek[2], and Sinem Bozkurt Keser[2]*

[1]Department of Software Engineering, Faculty of Engineering and Architecture, Eskişehir Osmangazi University, Eskişehir, Türkiye

[2]Department of Computer Engineering, Faculty of Engineering and Architecture, Eskişehir Osmangazi University, Eskişehir, Türkiye

Abstract

In the contemporary context of the obesity epidemic and its associated comorbidities, early detection of individuals at risk is critical. Artificial intelligence and machine learning techniques offer substantial potential for automating obesity risk assessment, enabling early diagnosis and intervention. However, the development of robust predictive models is often hampered by limited or imbalanced datasets. Synthetic data generation has emerged as a key solution, allowing the expansion and balancing of data while preserving privacy. Recent surveys highlight that the synthetic minority oversampling technique (SMOTE) is a leading method for data generation in obesity detection. In line with this, our study analyzed the Estimation of Obesity Levels dataset, a dataset from the University of California, Irvine repository, focused on dietary habits and physical condition, which suffers from class imbalance. We compared three synthetic data generation approaches: SMOTE—nominal and continuous, variational autoencoders, and conditional tabular generative adversarial network. We trained multiple classifiers on the generated datasets and evaluated their performance. Classifiers trained on data including height and weight (i.e., body mass index [BMI]-related features) achieved F1-scores of up to 98.16%, as expected due to the direct role of BMI in obesity classification. Crucially, models trained without height and weight still achieved an F1-score of 74.48% when synthetic augmentation was used, demonstrating that useful obesity prediction models can be developed even in the absence of explicit anthropometric measures. These results indicate that synthetic data can enable accurate classification when key features are missing or when data are scarce.

Keywords: Obesity; Synthetic data; Tabular data; Data augmentation; Machine learning; Class imbalance

***Corresponding author:**
Sinem Bozkurt Keser
(sbozkurt@ogu.edu.tr)

Citation: Eren HA, Emek Hİ, Keser SB. Comparison of synthetic data generation techniques for obesity level prediction based on dietary habits and physical status. *Artif Intell Health*. 2025;2(4):47-74. doi: 10.36922/AIH025140027

Received: April 1, 2025

Revised: June 2, 2025

Accepted: June 10, 2025

Published online: June 25, 2025

1. Introduction

According to the World Health Organization, obesity is defined as the accumulation of fat in the body to an extent that impairs health. The rise in obesity rates has become a growing concern not only in high-income countries but also in middle- and low-income nations.[1] The increasing prevalence of obesity across all age groups is linked to

health issues such as hypertension, diabetes, certain types of cancer, and musculoskeletal disorders.[2] Globalization has transformed obesity into a global public health challenge, demanding attention and coordinated action in the international policy arena.[3] Obesity is a major, yet preventable, global health condition, with a high and rising prevalence among children and adolescents, leading to serious health complications and substantial healthcare costs.[4] In light of these circumstances, early diagnosis of obesity becomes critically important. By analyzing individual characteristics, it is possible to predict an individual's risk of developing obesity.

Artificial intelligence (AI) methodologies hold great promise for automating obesity risk estimation by enabling early diagnosis and timely intervention.[5] For example, predictive models based on dietary and lifestyle features can identify individuals at elevated risk even before clinical obesity manifests. However, real-world datasets (especially survey-based ones) are often small, imbalanced, or contain missing values. Synthetic data generation provides a means to overcome these challenges by creating artificial records that replicate the statistical properties of real data.[6,7] This can improve model training and generalization without compromising patient privacy.[8]

Recent evidence underscores the multifactorial nature of obesity. For instance, early-life nutrition and feeding practices have long-term effects on weight trajectories: exclusive breastfeeding is associated with a lower risk of childhood overweight and obesity.[9] Studies of adult obesity also emphasize the role of dietary patterns and psychosocial factors. Sobas *et al.*[10] identified distinct dietary patterns ("prudent" healthy diet versus "fast food & alcohol") among bariatric surgery candidates, with the latter linked to more severe obesity.[10] Colonnello *et al.*[11] found that dysfunctional eating behaviors (e.g., night eating, food cravings) correlate with lipid and metabolic abnormalities in obese patients.[11] El-Sehrawy *et al.*[12] showed that a high triglyceride-glucose (TyG) index (a marker of insulin resistance) is associated with adverse lipid profiles and disordered eating in obesity.[12] Psychological stress is also implicated – Kuckuck *et al.*[13] demonstrated that long-term stress (measured by hair cortisone) is associated with hedonic eating tendencies in obese individuals.[13] Together, these studies highlight that beyond anthropometric measures, a combination of diet quality, metabolic markers, and behavioral patterns influence obesity outcomes. This study does not attempt to discover new causal factors; rather, it focuses on the methodological contribution of using synthetic data to improve obesity prediction models based on available dietary/behavioral features. Specifically, it examines the extent to which predictive accuracy can be maintained when key features (height and weight) are unavailable.

A review of the literature reveals that synthetic data generation is widely applied, with the synthetic minority oversampling technique (SMOTE) being one of the most commonly used approaches.[14-17] In line with this, the present study analyzes the Estimation of Obesity Levels Based on Eating Habits and Physical Status (EOL) dataset, which suffers from an unbalanced class distribution. To address the issue of limited sample size, various synthetic data generation techniques were employed.[18,19] Furthermore, an AI system was developed using machine learning (ML) algorithms to estimate obesity levels based on individuals' eating habits and physical status.

The performance of ML models trained on data generated using different techniques – namely variational autoencoders (VAE), generative adversarial network (GAN), and SMOTE – nominal and continuous (SMOTE-NC)—was compared. The next section of this paper presents a summary of related work, including the datasets used, methodologies applied, and results reported. The Materials and Methods section describes the dataset characteristics, synthetic data generation approaches, preprocessing procedures, and interrelationships among attributes. It also details the ML algorithms and evaluation methodology used to assess model effectiveness. The Results and Discussion section presents model outputs through various graphs and tables. Finally, the manuscript concludes with a summary and suggestions for future research.

2. Literature review

Numerous studies in the literature have addressed the problem of obesity detection, with particular emphasis on dataset construction and the development of ML models. Table 1 summarizes the key characteristics of the datasets used in these studies, the ML techniques applied, and the corresponding performance metrics reported.

Palechor *et al.*[14] developed a dataset for obesity level classification using data collected from individuals in Mexico, Peru, and Colombia. The dataset comprises 17 attributes related to eating habits and physical condition. Of the 2,111 instances, 23% were collected directly from users via a web platform, while the remaining 77% were synthetically generated using SMOTE. Classification of individuals was based on their body mass index (BMI) values. Subjects with a BMI below 18.5 were categorized as underweight, those with values between 18.5 and 24.9 as normal weight, and those between 25 and 29.9 as overweight. BMI values of 30 and above indicated obesity, which was further divided into three classes: 30–34.9 as

Table 1. Summary of literature on obesity risk prediction using machine learning (ML) algorithms

Study	Dataset	ML algorithm	Results
Helforoush and Sayyad[15]	UCI Obesity dataset (2,111 samples; 17 features)	ANN + PSO hybrid; compared with baseline regression	The ANN-PSO model achieved an accuracy of 92%, outperforming standard regression methods. SHAP analysis identified weight and height as the most influential features
Ayub *et al.*[16]	UCI Obesity dataset	Attention Bi-LSTM deep network	The proposed model achieved 96.5% accuracy in obesity classification, surpassing previous approaches. The integration of an attention mechanism enhanced the model's ability to capture feature influence
Shakti *et al.*[17]	UCI Obesity dataset	Multiple comparisons: k-NN, SVM, RF, GBM, MLP	The MLP achieved the highest accuracy of 97.2%, followed by GB with~96.2%. These results highlight the advantage of incorporating diverse features to improve classification performance
Yağmur[18]	UCI Obesity dataset	DT + POA (hybrid model)	The hybrid DT–POA model with fuzzy tuning outperformed the baseline DT, demonstrating improved classification performance for obesity levels
Özkurt[19]	UCI Obesity dataset	XGBoost, RF, NB, k-NN, DT (+ SHAP XAI)	XGBoost achieved the highest accuracy of 92%. SHAP analysis identified key predictors, including family history of obesity and vegetable intake
Wang[20]	UCI Obesity dataset (height/weight excluded)	Ordinal versus multinomial Logit; LogitBoost; SVM, NB, RF, k-NN	The LogitBoost model achieved the highest performance with~70% accuracy (Kappa=0.65). Other ML models yielded accuracies ranging from 75% to 79%. The overall lower accuracy was attributed to the exclusion of BMI-related features. Nonetheless, active transportation (e.g., biking), and family history were identified as key predictors
Okpe *et al.*[21]	UCI Obesity dataset	Multilayer perceptron ANN	A tuned ANN achieved 97% accuracy in multi-class obesity prediction, demonstrating that high accuracy can be attained with a relatively simple NN architecture
Azad *et al.*[22]	UCI Obesity dataset	Stacked ensemble (GBM, XGB, etc.) + LIME explanations	The stacking ensemble model achieved~98% accuracy, outperforming previous models (~97.8%). Model explainability was enhanced through the integration of LIME
Solomon *et al.*[23]	UCI Obesity dataset	Hybrid voting ensemble (XGBoost + GBM + MLP)	The ensemble model achieved an accuracy of 97.16%, surpassing the single XGBoost model (~96.4%). These results set a high benchmark for future studies in obesity prediction
Kaur *et al.*[24]	UCI Obesity dataset	GB, BME, XGBoost, RF, SVM, k-NN	XGBoost achieved 97.79% accuracy with a 70 – 30 train-test split, followed by GBM with~97.16%. The results demonstrated the superiority of ensemble methods. In addition, the model provided personalized diet recommendations based on predictive outcomes
Muliawan *et al.*[25]	Kaggle Obesity dataset (2,111 samples; 17 features)	RF	An accuracy of 81.76% was achieved using only eating habit parameters, validating the effectiveness of RF as a screening tool for obesity risk based solely on dietary data
Choudhuri *et al.*[26]	UCI Obesity dataset	Hybrid ML model (combining algorithms)	A hybrid approach was proposed for estimating obesity levels, combining multiple ML techniques. This method improved accuracy compared to individual models and has been cited in subsequent studies for its pioneering contribution
Cervantes and Palacio[27]	UCI Obesity dataset (original introduction)	Computational intelligence methods (e.g., ANN, fuzzy)	An early study achieved viable obesity level prediction, laying the groundwork for the application of ML on this dataset and serving as a baseline in later research
Ganie *et al.*[28]	Kaggle Obesity dataset (2,111 samples; 17 features)	Bagged DT, RF, extra tree, XGBoost, GB, CatBoost, voting classifier	The proposed model achieved 98.10% accuracy in obesity classification, outperforming previous approaches. The ensemble of boosting algorithms effectively captured complex patterns in lifestyle data
Nagarajan *et al.*[29]	UCI Obesity dataset	TabNet, XGBoost, RF, MLP, bagging, DT, SVM, k-NN, SGD, AdaBoost, stacking, GB	The proposed model achieved 99.3% accuracy in obesity classification, outperforming previous approaches. The use of SMOTE and deep learning techniques enhanced learning from imbalanced classes
Umoh *et al.*[30]	UCI Obesity dataset	KNN, SVM, bagging, stacking, voting, LR, DT, AdaBoost	The proposed model achieved 93.97% accuracy in obesity classification. Optimization through feature selection techniques improved the model's understanding of dietary and physical habits

(*Contd...*)

Table 1. (*Continued*)

Study	Dataset	ML algorithm	Results
Vairachilai *et al.*[31]	Kaggle COVID-19 Healthy Diet dataset	Protein Food Item Prediction Regression model	The proposed model achieved high predictive accuracy, with MAPE of 29% for meat and milk and 31% for oil crops and vegetable products. The integration of protein-rich food variables allowed refined modeling of feature influence in obesity prediction
Forte *et al.*[32]	FITescola® project dataset	CNN	The proposed model achieved 75% accuracy in obesity classification. The inclusion of physical fitness variables improved feature interpretability and overall model performance
Yağın *et al.*[33]	Physical Activity and Eating Habits dataset from İnönü University; includes alcohol use, device use, and meal frequency	Trained NN with Bayesian optimization	The proposed model achieved 93.06% accuracy in obesity classification, outperforming prior methods. The integration of Bayesian optimization enhanced the model's ability to select critical features
Gözükara Bağ *et al.*[34]	Web-based public dataset on physical activity and nutrition (gender, BMI, diet, etc.)	LR, RF, XGBoost with Bayesian optimization	The proposed model achieved 99.33% accuracy using logistic regression, with improved classification accuracy after feature selection. The inclusion of nutritional and activity data further strengthened the model's predictive capacity

Abbreviations: ANN: Artificial neural network; BME: Bagging meta-estimator; Bi-LSTM: Bidirectional long short-term memory; BMI: Body mass index; CNN: Convolutional neural network; COVID-19: Coronavirus disease 2019; DT: Decision tree; GB: Gradient boosting; GBM: Gradient boosting machine; k-NN: k-nearest neighbors; LIME: Local interpretable model-agnostic explanations; LogitBoost: Logistic regression boosting; LR: Logistic regression; MAPE: Mean absolute percentage error; MLP: Multi-layer perceptron; NB: Naïve Bayes; NN: Neural network; POA: Pelican optimization algorithm; PSO: Particle swarm optimization; RF: Random Forest; SGD: Stochastic Gradient Descent; SHAP: Shapley additive explanations; SVM: Support vector machine; UCI: University of California, Irvine; XAI: Explainable artificial intelligence; XGBoost: Extreme gradient boosting.

obesity type I, 35–39.9 as obesity type II, and 40 or higher as obesity type III.

In the study by Helforoush and Sayyad[15], titled *Hybrid Metaheuristic ANN-PSO*, various ML models were applied for obesity risk prediction. The authors proposed a hybrid artificial neural network optimized using particle swarm optimization (ANN-PSO). When evaluated on the University of California, Irvine (UCI) obesity dataset – which contains 2,111 records and 17 features related to dietary habits and physical conditions – the ANN-PSO model achieved an accuracy of ~92%, outperforming traditional regression models. To enhance interpretability, the study employed Shapley additive explanation analysis, which revealed that weight and height were among the most influential features in predicting obesity levels. These findings highlight the potential of metaheuristic optimization methods to improve the performance of neural networks in personalized obesity risk profiling.

Ayub *et al.*[16] developed an attention-enhanced bidirectional long short-term memory (ABi-LSTM) model to classify individuals into obesity categories using the same dataset. Their deep learning architecture incorporated an attention mechanism to emphasize key features – such as height, weight, and activity level – allowing the model to capture complex patterns within the data. The proposed ABi-LSTM achieved a multiclass classification accuracy of 96.5%, representing a substantial improvement in precision, recall, and F1-score over existing approaches. The authors highlighted this result as a paradigm shift, demonstrating the effectiveness of attention-based deep sequential models in enabling accurate obesity risk prediction.

Shakti *et al.*[17] evaluated multiple ML frameworks on the UCI obesity dataset, which contains 2,111 instances with 17 attributes related to eating habits and lifestyle factors. The models tested included k-nearest neighbors (k-NN), support vector machine (SVM), random forest (RF), gradient boosting (GB), and a multilayer perceptron (MLP) neural network. Among these, the MLP classifier achieved the highest accuracy at 97.2%, followed closely by GB at ~96.2%. These findings highlight that incorporating diverse features – such as dietary habits and physical activity – alongside robust learning algorithms like neural networks (NNs) can yield high classification performance. The study emphasizes that such levels of accuracy are essential for enabling targeted interventions for individuals at risk of obesity.

Yağmur[18] proposed a hybrid model that combines a decision tree (DT) classifier with the pelican optimization algorithm (POA), a metaheuristic optimization technique, to enhance obesity level classification. Utilizing the 2,111-instance dataset, the model applied fuzzy parameter tuning via POA to optimize the tree's decision thresholds for multiclass categorization. The hybrid DT-POA approach reportedly outperforms the standard DT model in predicting obesity levels. Although the precise accuracy value is not explicitly stated, the author highlights the

model's effectiveness and suggests that it can serve as a robust tool to assist healthcare professionals in obesity risk assessment. This study illustrates how evolutionary optimization algorithms can improve the performance of traditional classifiers in this domain.

Özkurt[19] implemented multiple ML algorithms in conjunction with explainability techniques to predict obesity risk. The study utilized data from 2,111 individuals in the UCI obesity dataset, which contains attributes related to dietary habits and physical conditions. A range of ML classifiers – including DT, RF, Naïve Bayes, k-NN, and extreme gradient boosting (XGBoost) – were evaluated. Among these, the XGBoost model achieved the highest classification accuracy at approximately 92% for obesity level prediction. To enhance interpretability, the author employed Shapley additive explanations to identify key features influencing the model's decisions. The analysis revealed that family history of obesity, vegetable intake, and frequency of between-meal consumption were among the most influential predictors. These findings demonstrate that boosting algorithms, when integrated with explainable AI (XAI) techniques, can deliver both high predictive performance and valuable insights into obesity-related risk factors.

In a related study, Wang[20] presented their findings in *E3S Web of Conferences*, focusing on obesity level prediction using lifestyle habit features while deliberately excluding direct anthropometric measures such as height and weight to assess model generalizability. The study evaluated a range of ML algorithms, including logistic regression variants (ordinal and multinomial), ensemble methods (LogitBoost and XGBoost), and standard classifiers (Naïve Bayes, SVM, RF, and k-NN). Among these, the LogitBoost ensemble achieved the highest performance, with an accuracy of ~70% and a Kappa statistic of ~0.65. In contrast, the XGBoost model performed poorly, reaching an accuracy of ≤20% due to the exclusion of key features. Other models, such as SVM, k-NN, and RF, achieved accuracies ranging from 75% to 79%. Although these values are lower than those reported in studies that incorporate BMI-related features, the author provided important insights. Specifically, they emphasized that when anthropometric data are unavailable, lifestyle indicators play a critical role in obesity prediction. Feature importance analysis revealed that the mode of transportation (e.g., riding a bike) was the most influential predictor, followed by family history of overweight and frequency of vegetable consumption. This comparative study suggests that even in the absence of direct body measurements, lifestyle-related attributes can still support reasonably accurate obesity risk assessments.

Okpe *et al.*[21] proposed a multilayer perceptron ANN model for multiclass obesity classification using the UCI obesity dataset. The study involved a comprehensive set of preprocessing steps, including handling missing values, encoding categorical attributes related to diet and physical activity, and additional data preparation procedures. A feedforward ANN was then implemented in Python and trained on the preprocessed dataset. The model achieved a classification accuracy of 97% across seven obesity categories, indicating its effectiveness in capturing the patterns between eating habits, physical conditions, and obesity outcomes. The authors emphasized that careful data cleaning and hyperparameter optimization were critical to achieving this high level of performance. Their findings highlight that even relatively simple NN architectures can yield accuracy comparable to more complex or ensemble-based models when properly optimized.

Azad *et al.*[22] proposed a stacking ensemble model that integrates XAI techniques for obesity risk classification, published in early 2025. In their study, the researchers combined multiple base classifiers within a stacked architecture and employed local interpretable model-agnostic explanations (LIME) to provide local interpretability. The model was evaluated on the standard obesity dataset, achieving an accuracy of ~98%, which slightly outperformed previously reported best-performing models such as GB and XGBoost (~97.8%). Beyond the improved predictive performance, the integration of LIME offered valuable insight into individual predictions, addressing the "black-box" issue. Comparative analysis demonstrated that the proposed approach outperformed all prior studies in terms of classification accuracy. This research highlights the effectiveness of combining diverse classifiers through ensembling and underscores the importance of incorporating XAI techniques to enhance model transparency, particularly in clinical decision-making contexts.

Solomon *et al.*[23] introduced a majority-voting ensemble model composed of GB, XGBoost, and an MLP NN to classify obesity levels. Utilizing the Latin American obesity dataset, their hybrid ensemble achieved an accuracy of 97.16%, surpassing the best-performing individual model (XGBoost), which attained 96.37%. This result, published in *Diagnostics* in 2023, established a high-performance benchmark and has since been frequently cited by 2024 studies as a state-of-the-art reference. By comparing multiple algorithms, the authors demonstrated that an ensemble model can effectively leverage the strengths of its individual components. The majority-voting approach outperformed all single classifiers, highlighting the advantage of combining diverse learning paradigms. The impact of this work is further reflected in subsequent research, such as that of Azad *et al.*,[22] which aimed to exceed the benchmark established by this study.

In a seminal study, Kaur *et al.*[24] investigated the application of ML algorithms for obesity risk prediction and meal planning. Using the UCI obesity dataset, the researchers applied six ML algorithms – GB, Bagging meta-estimator, XGBoost, RF, SVM, and k-NN – to predict adult obesity risk. The models were evaluated under various train-test split ratios (90/10, 80/20, 70/30, etc.), with ensemble methods consistently demonstrating superior performance. Notably, XGBoost achieved an accuracy of up to 97.79% at the 70:30 split, followed closely by GB at ~97.16%. In contrast, simpler models such as k-NN and SVM showed lower accuracy, ranging from 82% to 87%. The study also featured a diet recommendation component generated based on the model's predictions, demonstrating a practical integration of ML with personalized dietary guidance. This early work established the reliability of ML models – particularly boosting ensembles – in predicting obesity-related outcomes, with the reported accuracy of XGBoost (~97.8%) serving as a benchmark in subsequent literature.

Muliawan *et al.*[25] focused on leveraging only eating habit features for obesity risk prediction, employing an RF classifier. The study utilized an open-access version of the 17-feature obesity dataset obtained from Kaggle, placing emphasis on dietary variables (e.g., frequency of high-calorie food consumption and meal frequency) while deliberately minimizing reliance on physical measurements. The RF model achieved an accuracy of 81.76% in distinguishing between high-risk and low-risk individuals. Although this performance is lower than that of models incorporating both dietary and physical attributes, it underscores the critical role of physical features in achieving optimal predictive accuracy. Nonetheless, the findings demonstrate that food intake patterns alone can yield approximately 82% accuracy, emphasizing the potential of ML algorithms in healthcare-related applications. The authors conclude that RF can serve as an effective screening tool in scenarios where detailed anthropometric data are unavailable.

Choudhuri *et al.*[26] proposed a hybrid ML model for obesity level estimation, utilizing the UCI obesity dataset. While the paper does not report specific performance metrics, the term "hybrid" suggests a combination of classification and optimization techniques. Subsequent studies have cited this work as an early example of integrating multiple classifiers to enhance prediction accuracy. This study is considered foundational in the adoption of ensemble and hybrid approaches within obesity prediction research. It paved the way for later works – such as that of Helforoush and Sayyad[15] – which further developed and refined these strategies.

In a related vein, the study by Cervantes and Palacio,[27] published in *Informatics in Medicine Unlocked* in 2020, is notable as one of the earliest applications of ML algorithms to the "Obesity Levels" dataset. The researchers employed computational intelligence techniques – potentially including neural networks or fuzzy systems – to estimate obesity levels. This pioneering work catalyzed broader interest in the dataset, contributing to the establishment of baseline results and illustrating the feasibility of obesity classification through ML methods.

Ganie *et al.*[28] explored the efficacy of ensemble learning techniques for predicting obesity risk using a publicly available Kaggle dataset focused on lifestyle behaviors. The study applied various ensemble learning methods, including RF, extra trees, XGBoost, and CatBoost, using both bagging and boosting strategies. Among these, XGBoost delivered the highest performance, achieving an accuracy of 98.1% and an F1-score of 96.5%. The findings demonstrate the robustness of ensemble models, particularly boosting techniques, in deriving predictive insights from multi-dimensional lifestyle datasets.

Nagarajan *et al.*[29] performed a comparative analysis of several ML and deep learning models for predicting obesity levels using a real-world dataset with 17 features, including demographic and health-related variables. To improve model performance on imbalanced classes, the authors implemented SMOTE. The algorithms tested included TabNet, XGBoost, GB, MLP, and RF. The GB algorithm achieved the highest accuracy of 99.3%, with XGBoost and TabNet following closely at 99% and 98.4%, respectively, validating the effectiveness of ensemble and deep learning models in healthcare data analysis.

Umoh *et al.*[30] focused on optimizing various ML classifiers to estimate obesity levels from physical activity and dietary data obtained through structured surveys. The dataset underwent thorough preprocessing, including normalization and feature selection. The study evaluated a range of classifiers, including SVM, GB, DT, and others. Among them, GB emerged as the top-performing model, achieving an accuracy of 97.23%. This research highlighted the significance of integrating robust feature selection with classifier tuning for effective obesity level prediction.

Vairachilai *et al.*[31] applied the protein intake prediction and response (PIPR) ML model to analyze the impact of dietary behavior on obesity during the COVID-19 pandemic. The dataset included comprehensive lifestyle and nutritional behavior indicators. Multiple ensemble learning algorithms, such as RF and extra trees, were evaluated in the study. The PIPR model stood out with an accuracy of 96.7%, demonstrating its capability to capture nuanced relationships between protein intake and obesity risk and confirming the value of ensemble strategies in obesity prediction tasks.

Forte *et al.*[32] developed a deep learning-based NN model aimed at classifying obesity risks among Portuguese adolescents. The model used the FITescola® dataset, which includes information on physical fitness levels and BMI percentiles. Leveraging the power of deep learning, specifically convolutional NNs, the study aimed to improve the detection of obesity risk patterns in youth. The proposed model achieved a classification accuracy of 96.3%, showcasing the potential of deep NNs to support early intervention strategies in public health contexts.

Yağın *et al.*[33] proposed a Bayesian-optimized NN for the estimation of obesity levels using a dataset focused on lifestyle factors and eating habits obtained from the UCI ML Repository. The study utilized a feedforward deep NN whose hyperparameters were tuned via Bayesian optimization to maximize predictive accuracy. This optimization improved the network's ability to identify significant patterns in the data by fine-tuning parameters such as learning rate and hidden layers. The final model achieved an accuracy of 96.5%, outperforming earlier approaches and demonstrating the effectiveness of combining NNs with optimization strategies.

Gözükara Bağ *et al.*[34] introduced a predictive modeling approach that integrates physical activity and nutritional habit data for classifying obesity levels. They utilized a dataset comprising 2,111 records from the UCI ML Repository, which included variables such as gender, BMI, dietary patterns, and physical activity. The study employed ML algorithms, including RF, k-NN, and XGBoost. Feature scaling and selection techniques were applied to enhance model performance. The highest classification accuracy of 98.87% was achieved using the XGBoost algorithm, underscoring its superiority in handling complex lifestyle-related data for obesity classification.

Several works underscore the impact of diet and lifestyle features on obesity classification. For example, studies using the EOL dataset have identified that eating habits (e.g., frequency of high-calorie food intake, number of meals) and lifestyle choices (e.g., mode of transport, frequency of physical activity) significantly influence obesity level predictions. These findings are consistent with nutrition research showing that "prudent" diet patterns (rich in fruits and vegetables) are linked to lower obesity, whereas fast-food – heavy patterns correlate with higher adiposity.[10] Obesity is closely tied to metabolic syndrome markers. The TyG index study and investigations of oxytocin levels illustrate that blood biomarkers and hormonal factors are often elevated in obesity and associated with eating behaviors.[12,13]

In addition to SMOTE, various over-sampling techniques have been adapted for multiclass problems in medicine. Yang *et al.*[35] reviewed multiclass oversampling for imbalanced health datasets, noting an emerging trend toward hybrid methods combining SMOTE with other strategies.[35] While SMOTE-NC (used in our study) is a straightforward approach that interpolates minority-class samples in mixed-type data, more complex generators like GANs can capture non-linear feature dependencies. Synthetic tabular data in health often requires careful evaluation; we leverage standard classification metrics to assess model performance on generated data.[7]

Recent work on GANs and VAEs shows they can simulate realistic clinical datasets. For instance, standalone reports on conditional tabular GANs (CTGANs) or VAE variants demonstrate their success in reproducing distributions of complex clinical features.[6,7] However, empirical comparisons of these methods (VAE versus GAN versus traditional oversampling) in specific applications like obesity remain limited, which motivates our empirical study. In summary, while many studies have achieved high accuracy in obesity prediction using ensemble or deep learning models, they typically rely on the original data (often including BMI-related attributes).

3. Materials and methods

3.1. Dataset definition

This study utilized the dataset titled E*stimation of Obesity Levels Based on Eating Habits and Physical Condition*.[5] The data were collected from individuals in Mexico, Peru, and Colombia, encompassing information on dietary habits, physical conditions, and obesity levels. The dataset contains a total of 2,111 instances and 17 attributes. The first 498 instances were collected directly from users, while the remaining samples were synthetically generated by Palechor *et al.*[14] using SMOTE. All analyses and synthetic data generation in this study were conducted using the 498 user-collected samples. The features included are gender, age, height, weight, family history of obesity, frequent consumption of high-calorie foods, frequency of vegetable consumption, number of main meals, consumption of food between meals, smoking, daily water consumption, calorie tracking, frequency of physical activity, frequency of using technological devices, alcohol consumption, type of transportation used, and obesity level. It is important to note that the dataset contains no missing values. The gender distribution is shown in Figure 1, with 271 males (54.4%) and 227 females (45.6%), indicating a relatively balanced sample.

As illustrated in Figure 2, the data indicate a predominance of affirmative responses, with 300 individuals (60.2%) supporting the proposition and 198 individuals (39.8%) opposing it. The distribution reflects a clear majority in favor of the proposition.

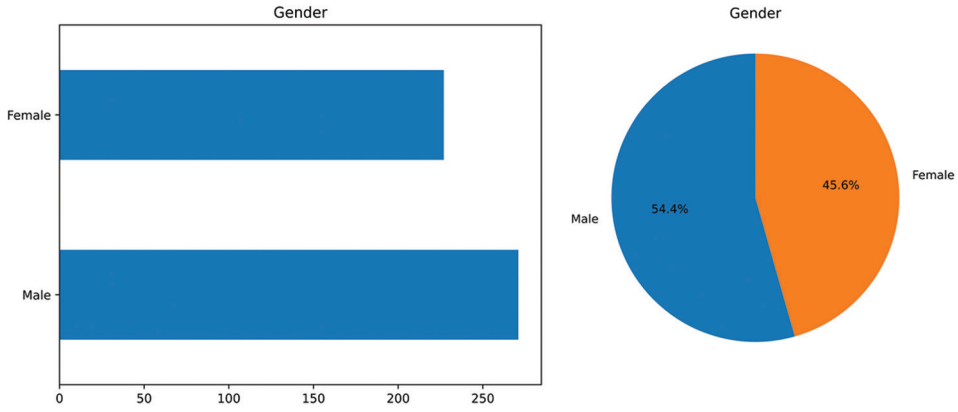

Figure 1. Gender distribution

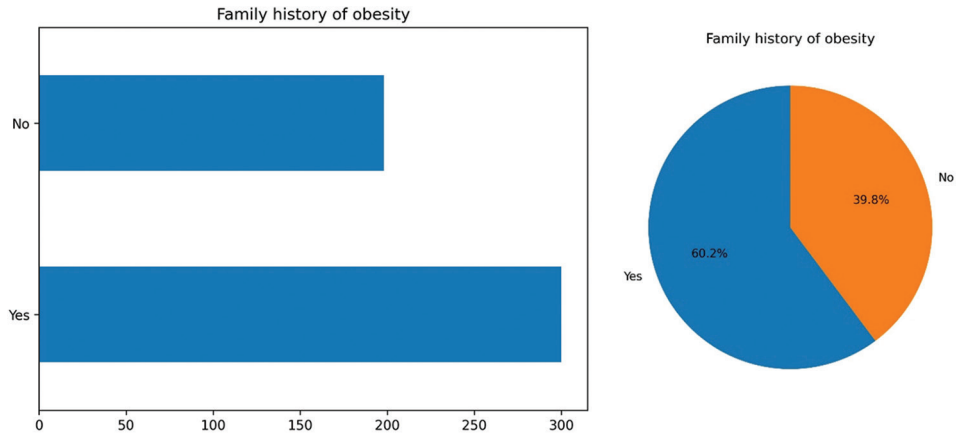

Figure 2. Distribution of family history of obesity

As shown in Figure 3, a similar pattern is observed in responses related to the consumption of high-calorie foods. The "yes" option predominates, with 348 individuals (69.9%) indicating frequent consumption, while the "no" option accounts for the remaining 30.1% of the sample.

The frequency of vegetable consumption was categorized into three response options: sometimes, always, and never. The number of responses for each category was 272, 194, and 32, respectively. As illustrated in Figure 4, these correspond to proportions of 54.6%, 39%, and 6.4%, respectively.

In the distribution of the number of main meals, the option of consuming three meals per day ranked first, with 344 individuals (69.1%). As illustrated in Figure 5, 108 individuals (21.7%) reported consuming only one meal, while 46 individuals (9.2%) reported consuming four meals daily.

As illustrated in Figure 6, the distribution of intermeal food consumption indicates that "sometimes" is the most frequently selected option, with 289 individuals. The remaining responses include "often" (n = 136), "always" (n = 53), and "no" (n = 20), each representing smaller portions of the sample.

The sample population consisted of 32 smokers and 466 nonsmokers, corresponding to 6.4% and 93.6% of the total, respectively. These proportions are illustrated in Figure 7.

An analysis of the daily water consumption reveals that 135 individuals consume <1 L, 266 individuals consume between 1 and 2 L, and 97 individuals consume more than 2 L/day. The corresponding scatter plot is presented in Figure 8.

A total of 55 individuals reported monitoring their caloric intake, whereas 443 individuals did not. As shown in Figure 9, these correspond to 11% and 89% of the sample, respectively.

The data indicate that 158 individuals (31.7%) engage in physical activity one or 2 days/week, while 162 individuals (32.5%) do not engage in any physical activity. In addition,

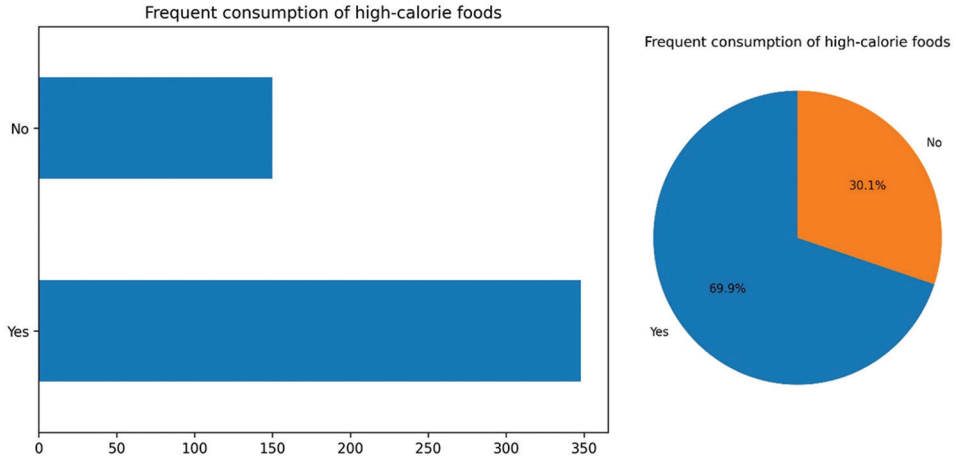

Figure 3. Distribution of frequent consumption of high-calorie foods

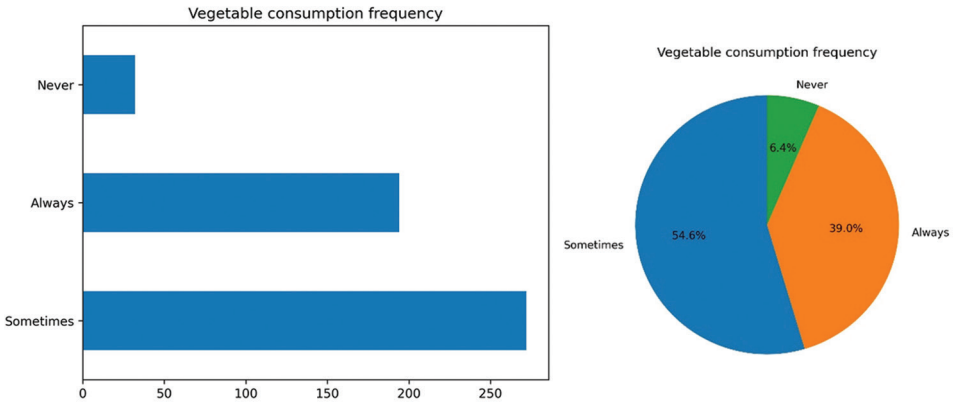

Figure 4. Distribution of frequency of vegetable consumption

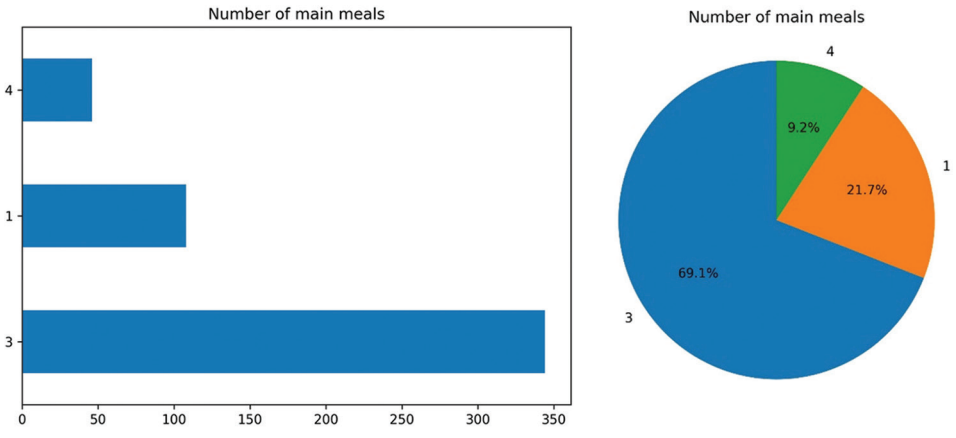

Figure 5. Distribution of the number of main meals

113 individuals (22.7%) are active for 2–4 days, and the remaining 65 individuals (13.1%) participate in physical activity for 4–5 days/week. The distribution of these values is illustrated in Figure 10.

A total of 243 individuals (48.8%) reported using technological devices for 0–2 h/day, whereas 181 individuals (36.3%) indicated daily usage of 3–5 h. As illustrated in Figure 11, the remaining 74 individuals

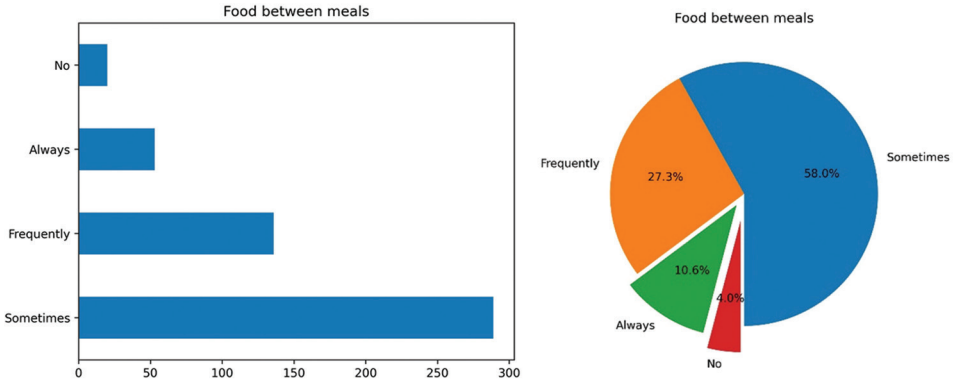

Figure 6. Distribution of food consumption between meals

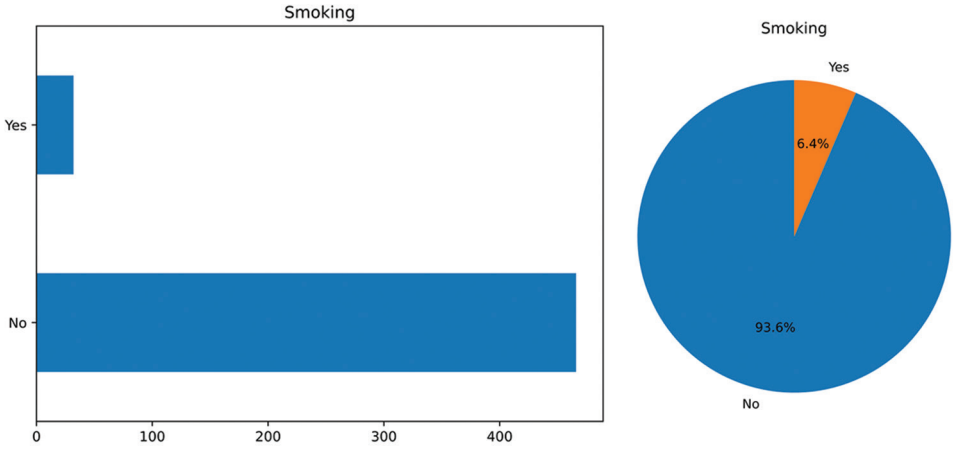

Figure 7. Distribution of smoking

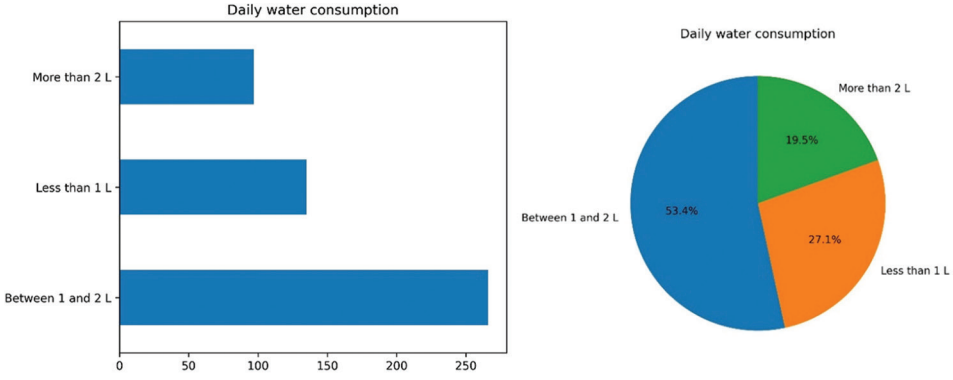

Figure 8. Distribution of daily water consumption

(14.9%) reported using technological devices for more than 5 h/day.

As illustrated in Figure 12, 273 individuals (54.8%) reported occasional alcohol consumption, while 179 individuals (35.9%) stated that they never consumed alcohol. In addition, 45 individuals (9%) reported frequent alcohol use, and one participant (0.2%) indicated consistent daily alcohol consumption.

Regarding transportation preferences, the data indicate a predominant reliance on public transportation, with 326 individuals (65.5%) selecting this option. Automobile use ranks second, preferred by 99 individuals (19.9%).

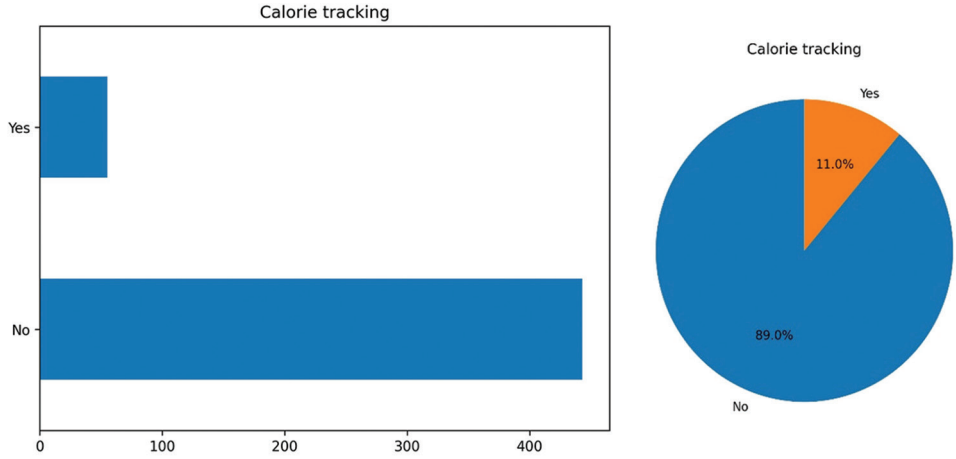

Figure 9. Distribution of calorie tracking

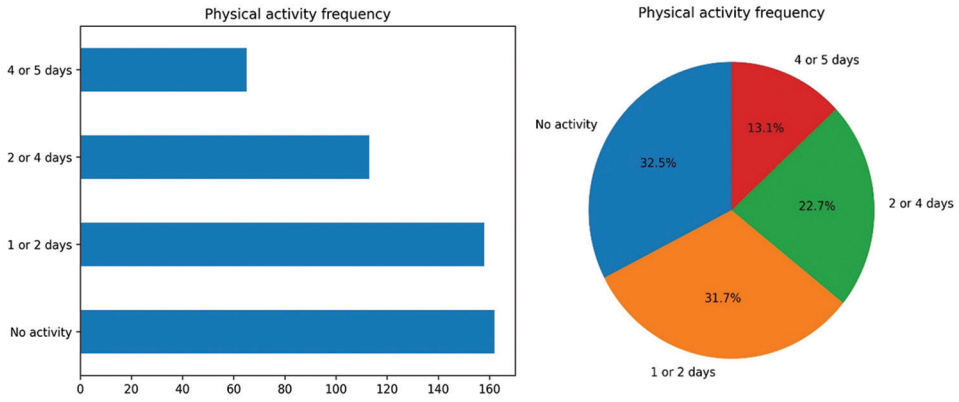

Figure 10. Distribution of physical activity frequency

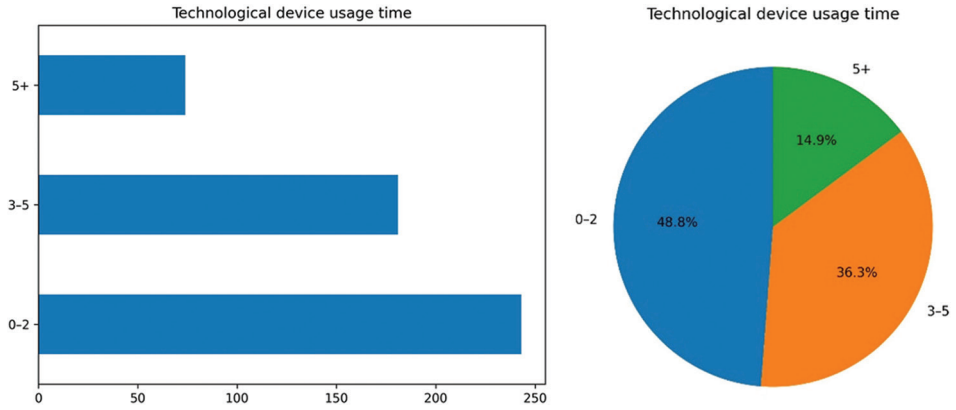

Figure 11. Distribution of duration of technological device use

Walking was chosen by 55 individuals (11%), followed by motorcycles (11 individuals; 2.2%) and bicycles (seven individuals; 1.4%). The distribution of these preferences is illustrated in Figure 13.

The original version of the dataset includes seven obesity level categories as class labels: underweight, normal weight, level I overweight, level II overweight, type I obese, type II obese, and type III obese. Due to class imbalance,

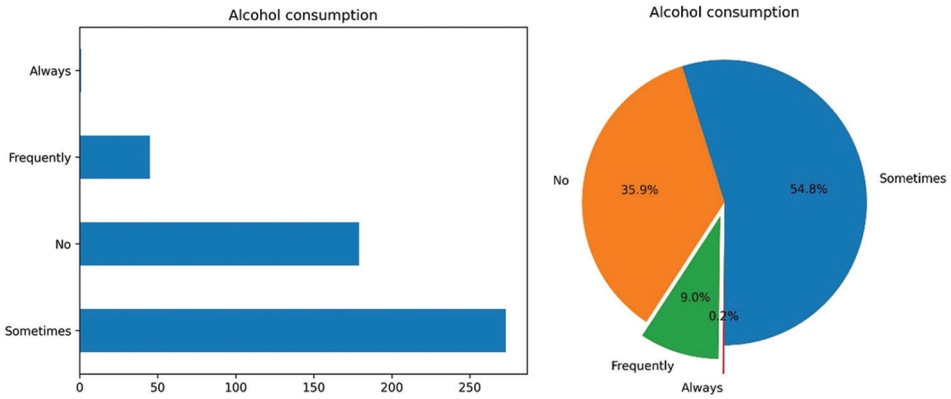

Figure 12. Distribution of alcohol consumption

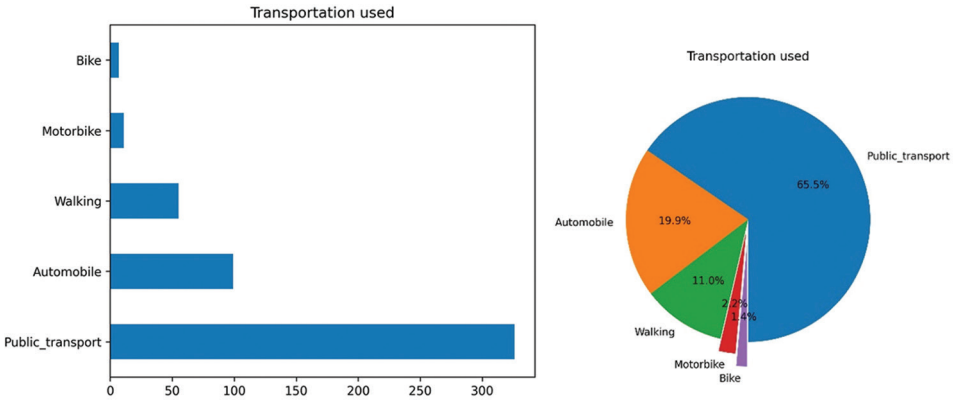

Figure 13. Distribution of transportation modes used

the overweight and obese subcategories were consolidated into single groups, resulting in four final categories. The distribution of the obesity level attribute after this merging is illustrated in Figure 14. According to the revised distribution, 37 individuals (7.4%) were classified as underweight, 284 individuals (57%) as normal weight, 116 individuals (23.3%) as overweight, and 61 individuals (12.2%) as obese.

The age values in the dataset ranged from 14 to 61 years. The distribution of ages is presented in Figure 15. The mean age was 23.15 years, with a standard deviation of 6.72, whereas the median age was 21.

The height values in the dataset ranged from 1.45 to 1.98 m. The mean height was 1.69 ± 0.09 meters, and the median was 1.68 meters. The distribution of height values is illustrated in Figure 16.

The dataset includes weight values ranging from 39 to 173 kg. The mean weight was 69.57 ± 17.01 kg, and the median was 67 kg. The distribution of weight values is depicted in Figure 17.

The mean weight within the underweight category was 53.8 kg for men and 46 kg for women. For individuals in the normal weight category, the mean values were 67.9 kg (men) and 56.5 kg (women). In the overweight group, mean weights were 82.7 kg for men and 71.4 kg for women. Finally, in the obese category, which represents the highest weight class, the averages increased to 106.4 kg for men and 86.3 kg for women. The relationship between obesity levels and weight by gender is illustrated in Figure 18.

As illustrated in Figure 19, individuals who did not consume vegetables and those who avoided eating between meals were predominantly in the younger age group. In addition, data suggest that underweight individuals and those who tracked their caloric intake also tend to be younger.

As illustrated in Figure 20, weight distributions are presented across the different obesity levels. As expected, the categories appeared in a sequential manner, with weight increasing progressively from lower to higher obesity classes.

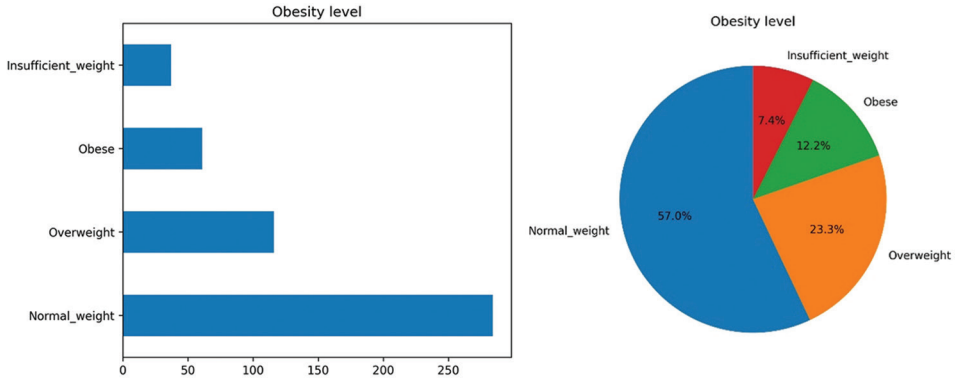

Figure 14. Distribution of obesity level

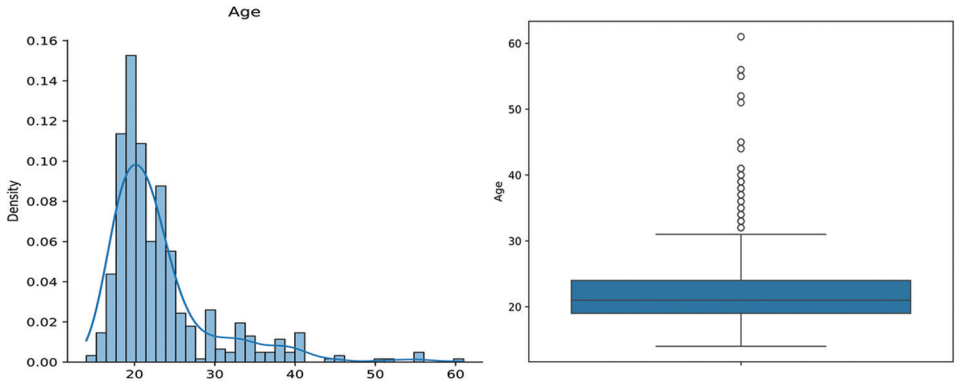

Figure 15. Age distribution

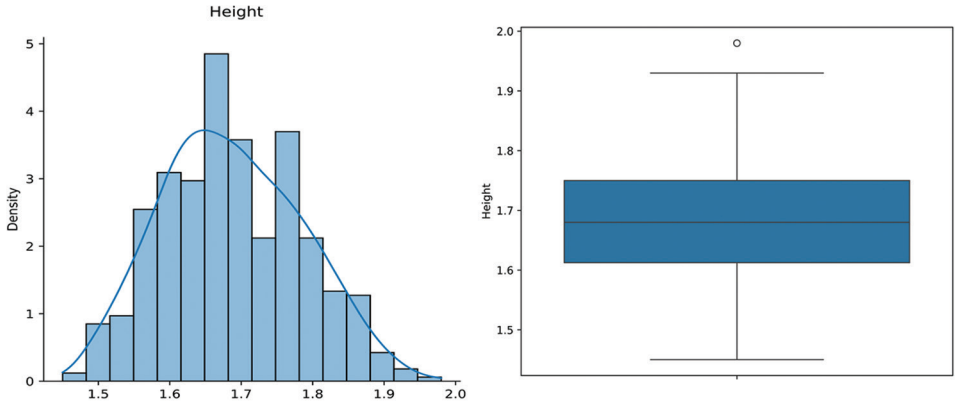

Figure 16. Height distribution

As illustrated in Figure 21, the axes of the graph represent height and weight, with color coding used to indicate different obesity levels. It is evident that, for a given height, the obesity class increased with weight, and male subjects tended to fall into higher obesity categories compared to female subjects. As BMI – the metric used to assign class labels – is directly correlated with height and weight (as shown in Equation I and Figure 21), the classes appeared to be linearly separable in the graph. However, including these attributes in model training may lead to an overestimation of performance. To illustrate this potential discrepancy, two distinct datasets were used: one that includes height and weight, and another that excludes them.

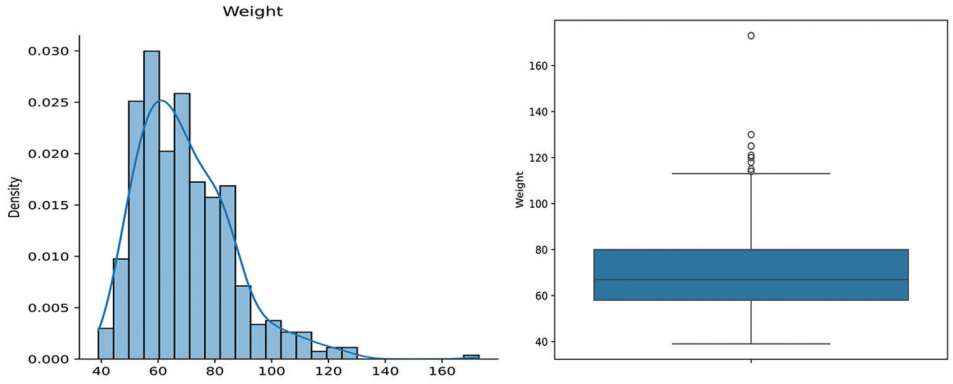

Figure 17. Weight distribution

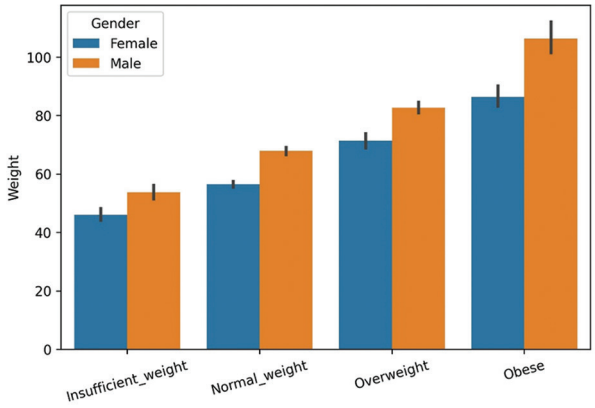

Figure 18. Associations of different levels of obesity with weight by gender

$$\text{BMI} = \frac{\text{Weight (in kg)}}{\text{Height}^2\text{ (in m)}} \quad \text{(I)}$$

As illustrated in Figure 22, the rows represent gender, the columns indicate whether individuals tracked their caloric intake, the axes correspond to age and weight, and the colors denote obesity classes. The figure reveals that individuals with higher levels of obesity were predominantly those who did not track calories and exhibited higher weight values. Furthermore, the data suggest that individuals who engaged in calorie tracking tended to be younger.

3.2. Synthetic data generation

The synthetic data generation methods employed in this study included the SMOTE-NC method from the *Imbalanced-learn* library by Lemaître *et al.*[36] and the VAE-based tabular VAE (TVAE) and GAN-based CTGAN by Xu *et al.*[37] methods in the Synthetic Data Vault (SDV) library by Patki *et al.*[38] Given that the majority class in the original dataset consisted of individuals with normal weight (284 samples), synthetic data were generated to match this sample size in each of the minority classes. After data generation, the final dataset comprised 1,136 instances, with equal representation across the four classes: underweight, normal weight, overweight, and obese.

SMOTE-NC is a variant of the SMOTE designed to address class imbalance by generating synthetic samples through interpolation. Unlike the original SMOTE algorithm, SMOTE-NC is capable of handling both numerical and categorical features, thereby producing synthetic data that more accurately represents the underlying structure of the original dataset. This method improves the diversity and representativeness of the minority class, ultimately contributing to more robust and generalizable model training.[39]

The TVAE is a generative model based on the VAE architecture, specifically designed to handle the heterogeneous nature of tabular data, which often includes a mix of continuous and categorical variables. The model consists of an encoder network that maps the input data into a latent space represented by Gaussian distributions and a decoder network that reconstructs the data from these latent representations. This structure enables TVAE to learn complex data distributions and supports conditional data generation by allowing specific attributes to be fixed during the sampling process. Once trained, TVAE can generate realistic synthetic tabular data by sampling from the latent space, providing a robust framework for addressing class imbalance and performing data augmentation tasks.[40]

The CTGAN extends the traditional GAN architecture by introducing modifications tailored to the unique characteristics of tabular data. While standard GANs – originally developed for image generation – struggle with the heterogeneity of tabular datasets, particularly due to mixed data types and the presence of discrete variables, CTGAN effectively addresses these limitations.

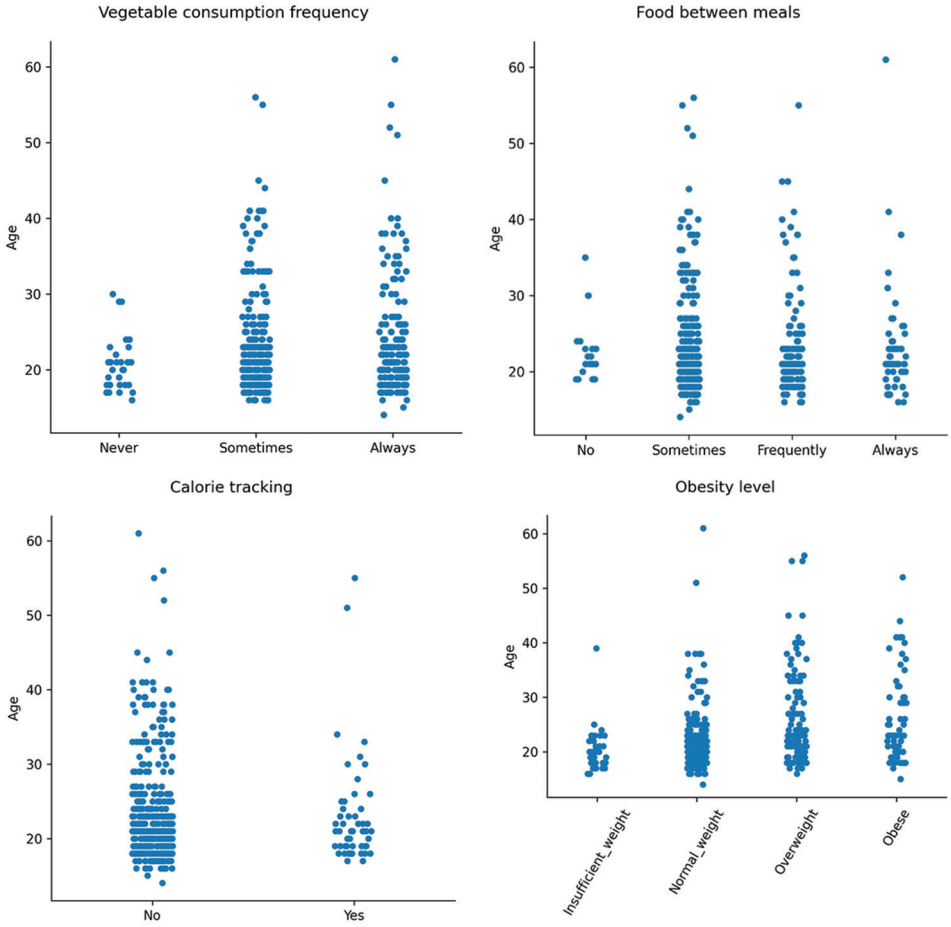

Figure 19. Frequency of vegetable consumption, food consumption between meals, calorie tracking, and obesity level by age

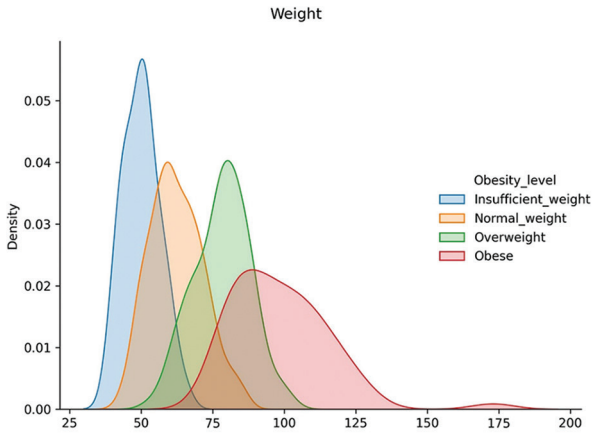

Figure 20. Weight distribution across obesity levels

It incorporates conditional data generation and mode-specific normalization techniques to model complex relationships in tabular data more accurately.[41]

The synthetic data generated using SMOTE-NC, which is designed to handle both numerical and categorical variables, includes numerical attributes such as age, height, weight, and number of main meals. These values were initially represented with up to 16 digits after the decimal point. Therefore, appropriate rounding procedures were applied to enhance data consistency. Specifically, the age and number of main meals were rounded to whole numbers, height to two decimal places, and weight to one decimal place. In contrast, no such adjustments were necessary for the synthetic data generated by VAE and GAN-based methods, as this issue did not occur. However, for these NN-based approaches, a separate model was trained for each class. It was observed that training a single model using all class samples resulted in lower performance, highlighting the advantage of class-specific model training in these architectures.

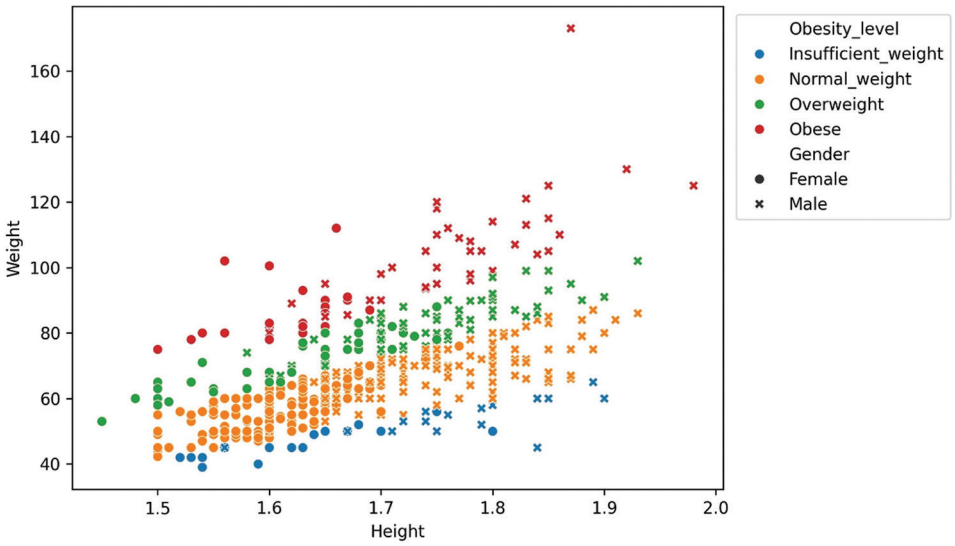

Figure 21. Obesity class distributions by gender on height and weight axes

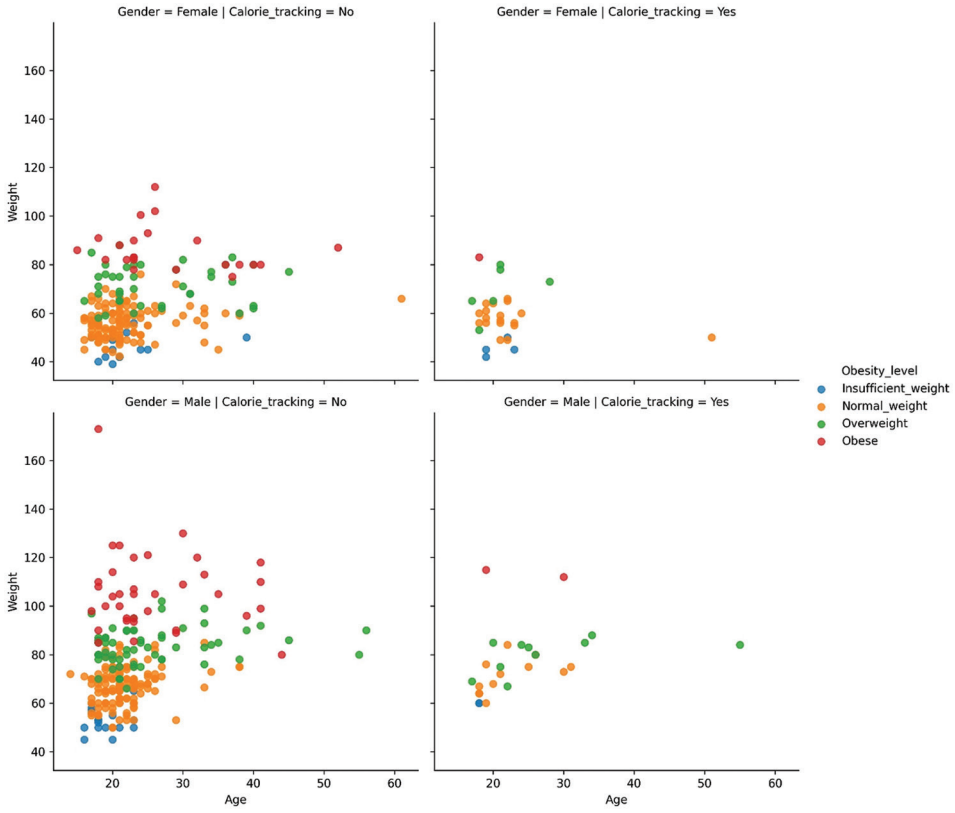

Figure 22. Associations between gender, calorie tracking, and obesity level on the age and weight axes

3.3. Data preprocessing and ML algorithms

Given that the majority of the attributes in the dataset are categorical, both one-hot encoding and label encoding methods were applied to prepare the data for input into ML algorithms. Among the attributes subjected to one-hot encoding, the gender variable was coded as male = 0 and female = 1. For binary attributes with "yes"/"no" responses, such as family history of obesity, frequent consumption of high-calorie foods, smoking,

and calorie tracking, "no" responses were encoded as 0 and "yes" responses as 1.

Label encoding was applied to attributes that exhibit an ordinal relationship among their values. These attributes include frequency of vegetable consumption, food consumption between meals, daily water consumption, frequency of physical activity, duration of technological device use, alcohol consumption, transportation preference, and obesity level (class label).

For frequency of vegetable consumption, the categories "never," "sometimes," and "always" were encoded as 0, 1, and 2, respectively. The responses "no," "sometimes," "often," and "always" for food consumption between meals were encoded as 0, 1, 2, and 3. Daily water consumption categories – "less than 1 liter," "between 1 and 2 liters," and "more than 2 liters" – were encoded as 1, 2, and 3. For frequency of physical activity, the categories "no activity," "1–2 days," "2–4 days," and "4–5 days" were assigned the values 0, 1, 2, and 3, respectively. Duration of technological device usage was categorized as "0 – 2 h," "3 – 5 h," and "5+ h" and encoded as 0, 1, and 2. Alcohol consumption levels ("no," "sometimes," "often," and "always") were encoded as 0, 1, 2, and 3.

For transportation preference, two different encoding strategies were tested. In one version, "public transportation" was assigned the base value 0, whereas in the other version, "walking" was assigned 0. The results of ML models showed no significant difference between these two approaches. As a result, the final version of the dataset adopted the encoding that prioritized "walking," with values ranging from 0 to 4. The obesity level, designated as the class label, was encoded from 0 to 3, where 0 represents the lowest level (underweight), and 3 represents the highest level (obese).

Following the encoding process, correlation heatmaps were generated for each of the datasets created using SMOTE-NC, TVAE, and CTGAN. These heatmaps were used to visualize the relationships between the attributes, where values close to 1 indicate strong positive correlations and values close to – 1 indicate strong negative correlations.

The correlation heatmap generated from the dataset synthesized using the SMOTE-NC method, presented in Figure 23, reveals notable relationships among several anthropometric variables. Significant correlations were observed between height and gender, weight and gender, weight and height, family history of obesity and weight, obesity level and weight, as well as obesity level and family history of obesity. The strongest correlation, with a

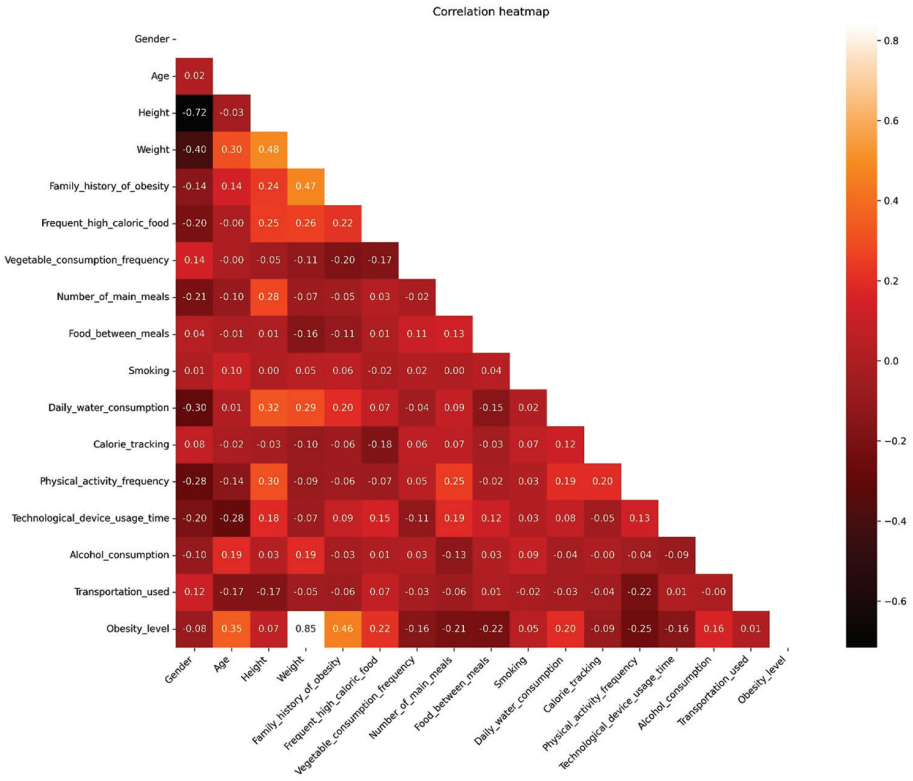

Figure 23. Correlation heatmap for the dataset generated using the synthetic minority oversampling technique—nominal and continuous

coefficient of 0.85, was identified between obesity level and weight, indicating a strong positive relationship.

As illustrated in Figure 24, the correlation heatmap generated from the dataset synthesized using the TVAE method reveals significant associations among various health-related metrics. These included the correlations between height and gender, family history of obesity and weight, obesity level and weight, obesity level and family history of obesity, and obesity level and frequency of physical activity. The strongest correlation was observed between obesity level and weight, with a coefficient of 0.90, indicating a very strong positive relationship.

As shown in Figure 25, the correlation heatmap generated from the dataset synthesized using the CTGAN method reveals notable associations between height and gender, height and weight, and obesity level and weight. A correlation coefficient of 0.84 was observed between obesity level and weight, indicating a strong positive relationship. The consistently high correlation between obesity level and weight across all three datasets (SMOTE-NC, TVAE, CTGAN) can be attributed to the direct role of weight in the calculation of BMI, which serves as the basis for obesity classification, as shown in Equation I and Figure 21.

In addition to encoding, the dataset was standardized using the StandardScaler function from the Scikit-learn library. For each ML algorithm, the training and testing process was repeated 100 times. During each iteration, models were evaluated using multiclass classification metrics: accuracy, precision, recall (sensitivity), and F1-score. These metrics were macro-averaged across classes. The performance metrics used in the study are defined in Table 2.

All metrics were computed using the actual (true) class labels and model predictions on the test set; a "correct prediction" means the predicted class matches the true label. In every run, the random_state parameter was set to values ranging from 0 to 99, based on the current iteration index. Stratified splitting was employed to divide the dataset into training and test sets while preserving class distribution. All classifiers available in the Scikit-learn library were evaluated, and the results of the five models with the highest F1 scores were reported.

4. Results and discussion

This section presents the performance metrics of the models trained on datasets generated using SMOTE-NC, TVAE, and CTGAN – the synthetic data generation techniques

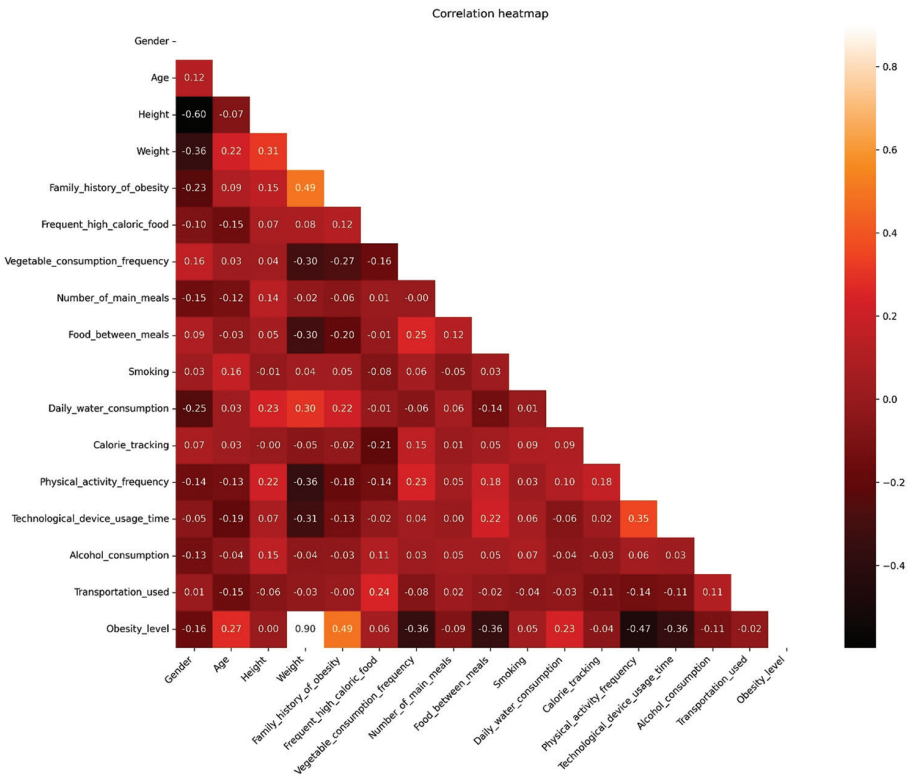

Figure 24. Correlation heatmap for the dataset generated using the tabular variational autoencoder

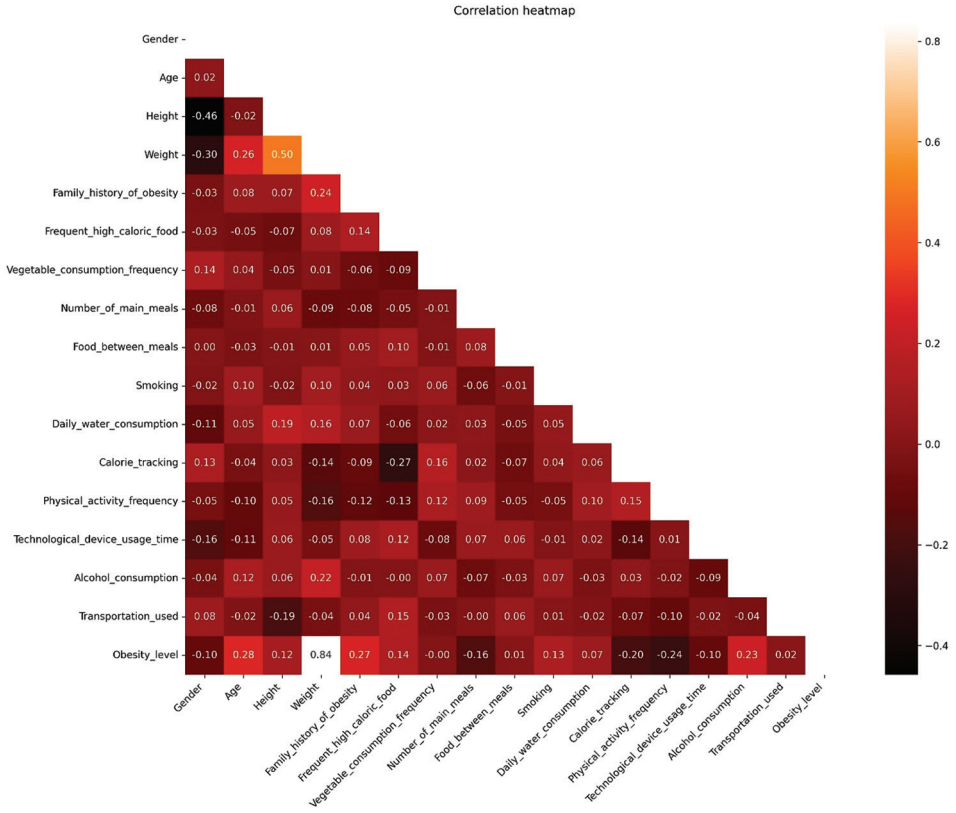

Figure 25. Correlation heatmap for the dataset generated using the conditional tabular generative adversarial network

Table 2. Performance metrics used in model evaluation

Metric	Explanation	Formula
Accuracy (ACC)	Gives the correct prediction rate of the model across all classes	$ACC=\frac{TP+TN}{TP+TN+FP+FN}$
Precision (PRE)	Shows how many positive predictions are actually positive	$PRE=\frac{TP}{TP+FP}$
Recall (REC)	Shows how many true positives are correctly predicted	$REC=\frac{TP}{TP+FN}$
F1-score	Is the harmonic mean of the accuracy and recall metrics	$F1=2x\frac{Precision\ x\ Recall}{Precision\ +\ Recall}$

Notes: TP: True positive; a positive sample correctly predicted as positive. TN: True negative; a negative sample correctly predicted as negative. FP: False positive; a negative sample incorrectly predicted as positive. FN: False negative; a positive sample incorrectly predicted as negative.

employed in this study. Each reported value represents the average performance of 100 independently trained models, utilizing all available classification algorithms in the Scikit-learn library. The results reflect the top five classifiers in terms of F1-score. Model performance is reported for two scenarios: one excluding the height and weight attributes, and one including them. As shown in Table 3 and Figure 26, the classifiers trained on the SMOTE-NC-generated dataset without height and weight information achieved average performance scores ranging from 70% to 75%.

When height and weight attributes were included, as shown in Table 4 and Figure 27, the average performance increased significantly, with F1 scores reaching up to 98.16%.

As illustrated in Table 5 and Figure 28, the dataset generated using the TVAE method yielded an average performance between 71% and 73% when height and weight attributes were excluded.

MODELS trained on the TVAE-generated dataset that included height and weight features achieved an F1 score of 97.49%. A comprehensive summary of these results is presented in Table 6 and Figure 29.

In the case of the dataset generated using CTGAN – the final synthetic data generation technique – classification models achieved lower performance compared to the other two methods when height and weight attributes were

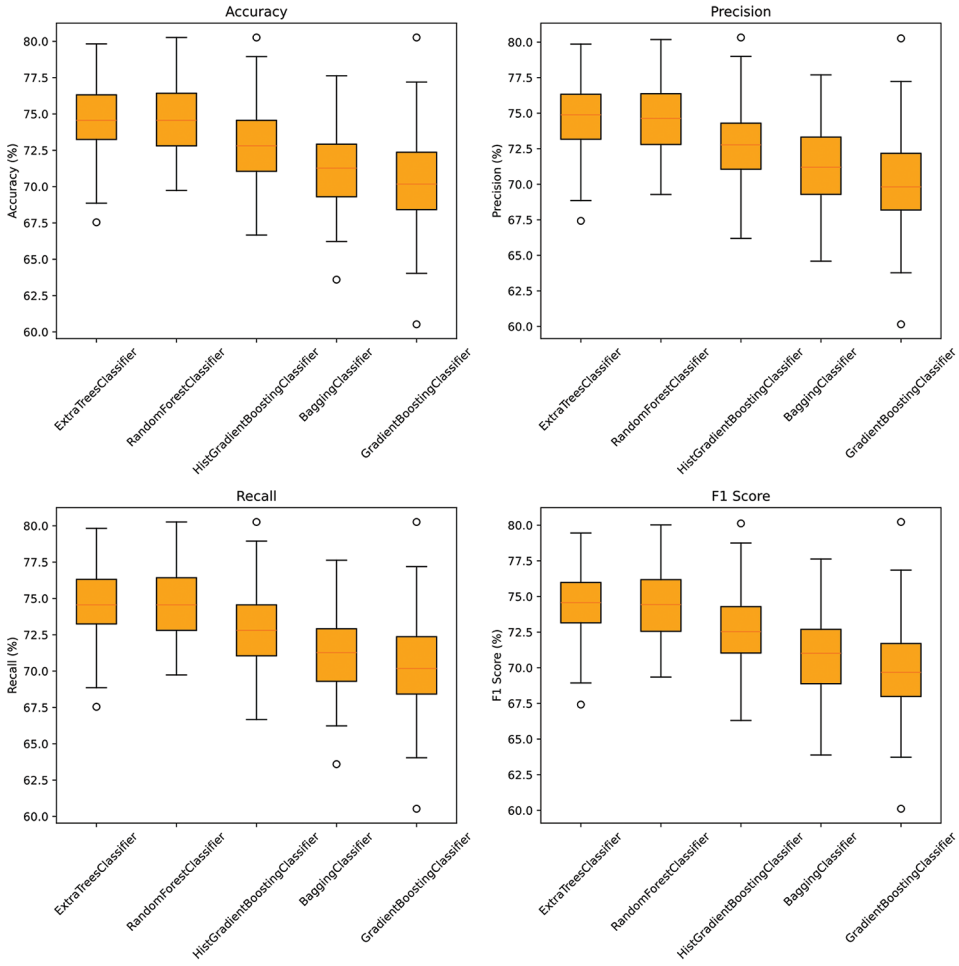

Figure 26. Performance metrics plots of the five most successful classifiers on the SMOTE-NC dataset (excluding height and weight attributes)

Table 3. Average performance metrics of the five most successful classifiers on the synthetic minority oversampling technique – nominal and continuous dataset (excluding height and weight attributes)

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ExtraTrees	74.62	74.72	74.62	74.48
RandomForest	74.71	74.62	74.71	74.45
HistGradBoosting	72.87	72.89	72.87	72.70
Bagging	71.24	71.27	71.24	70.98
GradBoosting	70.31	70.19	70.31	69.91

Table 4. Average performance metrics of the five most successful classifiers (using height and weight attributes) on the synthetic minority oversampling technique – nominal and continuous dataset

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LogisticRegCV	98.17	98.21	98.17	98.17
HistGradBoosting	96.61	96.65	96.61	96.61
GradBoosting	95.73	95.79	95.73	95.73
Bagging	94.55	94.64	94.55	94.55
LogisticReg	92.86	93.03	92.86	92.86

excluded. As shown in Table 7 and Figure 30, the models trained on this dataset reached an average F1 score of approximately 60%.

When height and weight were incorporated as input features, the performance of classifiers trained on CTGAN-generated data became comparable to those trained on data synthesized by other methods. A detailed analysis of Table 8 and Figure 31 showed that F1 scores range between 94% and 97%.

The strong performance of classifiers when using height and weight reaffirms known biology principles: BMI strongly separates obesity classes. A novel insight

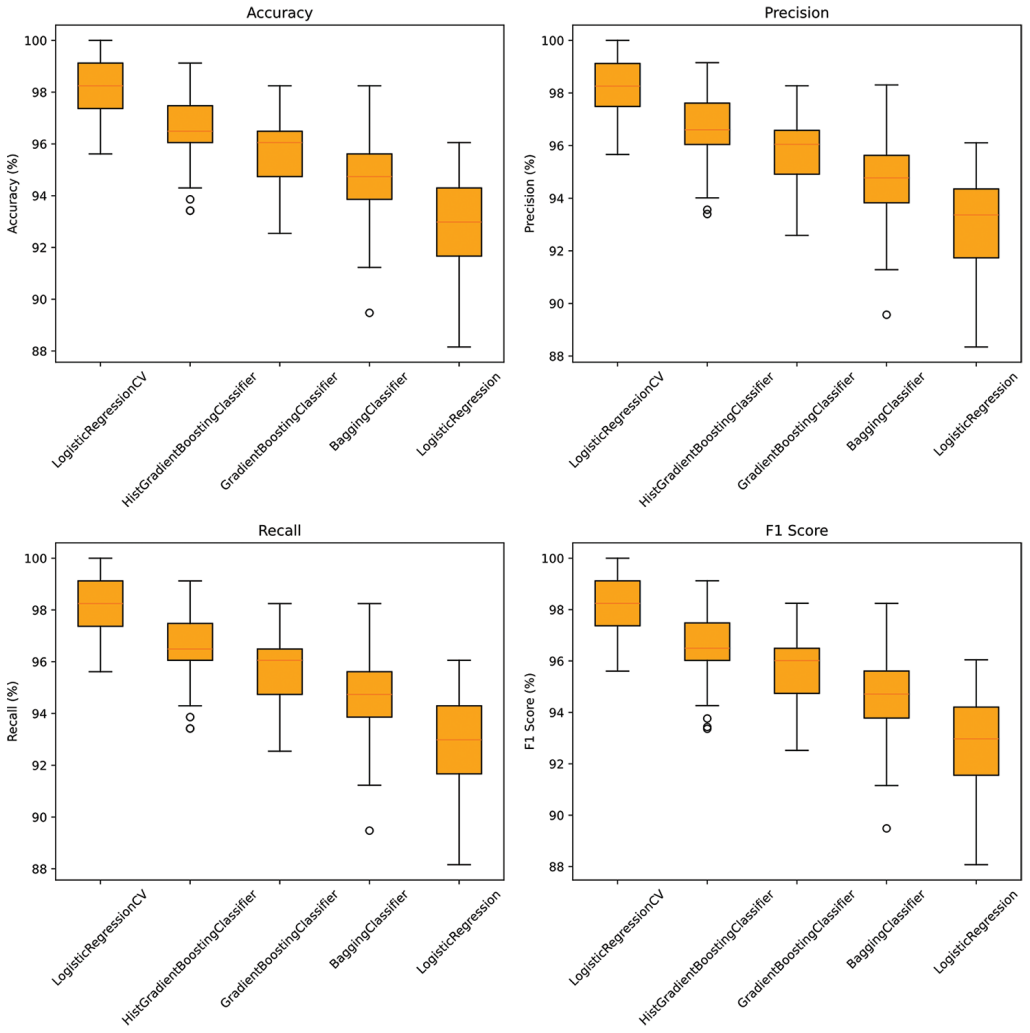

Figure 27. Performance metrics plots of the five most successful classifiers (using height and weight attributes) on the synthetic minority oversampling technique—nominal and continuous dataset

Table 5. Average performance metrics of the five most successful classifiers on the tabular variational autoencoder dataset (excluding height and weight attributes)

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVC	73.02	74.16	73.02	72.68
NuSVC	72.53	74.32	72.53	72.25
RandomForest	72.12	72.53	72.12	71.77
GradBoosting	71.51	71.31	71.51	71.12
ExtraTrees	71.31	71.66	71.31	71.09

Table 6. Average performance metrics of the five most successful classifiers on the tabular variational autoencoder dataset (using height and weight attributes)

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LogisticRegCV	97.49	97.54	97.49	97.49
HistGradBoosting	96.08	96.13	96.08	96.07
GradBoosting	94.54	94.59	94.54	94.52
Bagging	94.36	94.44	94.36	94.35
DecisionTree	92.75	92.87	92.75	92.74

from our study is that even without those direct measures, reliable classification (~75% F1 score) is possible by leveraging diet and lifestyle features through synthetic data augmentation. This finding is clinically relevant, where in many settings (e.g., telehealth surveys, electronic records lacking anthropometric data, or privacy-preserved

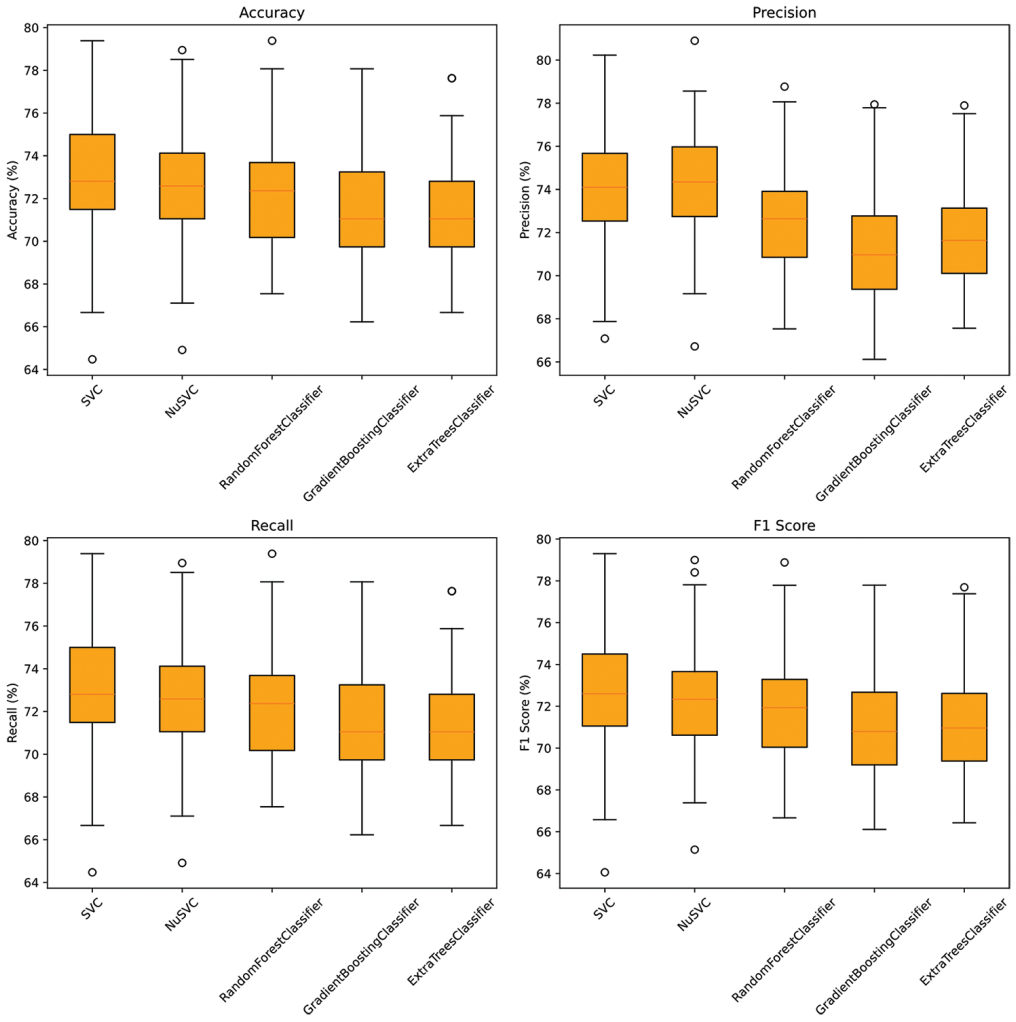

Figure 28. Performance metrics plots of the five most successful classifiers on the tabular variational autoencoder dataset (excluding height and weight attributes)

Table 7. Average performance metrics of the five most successful classifiers on the conditional tabular generative adversarial network dataset (excluding height and weight attributes)

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
GradBoosting	60.66	60.88	60.66	60.59
HistGradBoosting	59.53	59.80	59.53	59.49
RandomForest	59.25	59.03	59.25	58.92
ExtraTrees	57.40	57.34	57.40	57.19
Bagging	55.70	55.77	55.70	55.43

Table 8. Average performance metrics of the five most successful classifiers on the conditional tabular generative adversarial network dataset (using height and weight attributes)

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LogisticRegCV	97.45	97.50	97.45	97.45
HistGradBoosting	96.09	96.16	69.09	96.09
Bagging	95.53	95.64	95.53	95.53
GradBoosting	95.24	95.34	95.24	95.25
DecisionTree	94.26	94.35	94.26	94.25

research databases), height and weight may be unavailable or missing. Our results suggest that in such cases, synthetic data methods can help build models that still identify obesity risk with reasonable accuracy. This extends the known correlation of diet/behavior with obesity. For example, higher consumption of fast foods and irregular

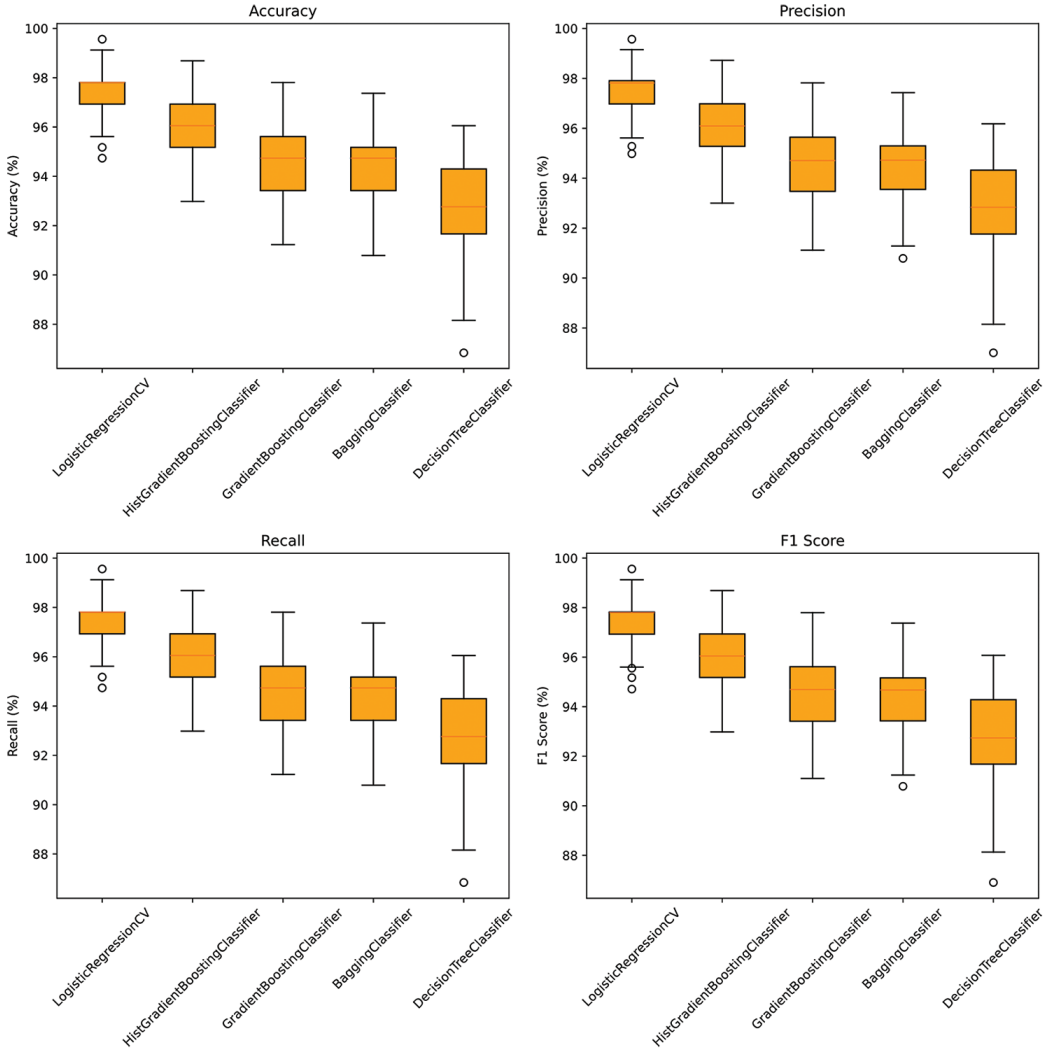

Figure 29. Performance metrics plots of the five most successful classifiers on the tabular variational autoencoder dataset (using height and weight attributes)

meal patterns are associated with higher obesity risk, and our synthetic augmentation appeared to capture these signals effectively for the ML models.

Moreover, these findings align with recent nutrition research. Colonnello *et al.*[11] found that dysfunctional eating behaviors (e.g., night eating) are correlated with lipid and metabolic abnormalities; we note that such behaviors are indirectly represented in our features (e.g., meal frequency, alcohol use).[11] El-Sehrawy *et al.*[12] reported that elevated TyG index values and disordered eating often co-occur in individuals with obesity, suggesting metabolic–diet linkages.[12] In our models, features related to eating patterns (e.g., intake of high-calorie foods, frequency of snacks) contribute to predictions, which is consistent with these clinical findings.

Our comparison highlights practical considerations for applying generative data methods in health. Consistent with Hernadez *et al.*,[7] we found that SMOTE-type oversampling and VAE-based generation can effectively balance and expand tabular health data.[7] The poorer performance of CTGAN (in the no-BMI case) suggests that GAN-based approaches may require more data or tuning to capture complex categorical relationships in this dataset. Importantly, synthetic data offer benefits beyond model accuracy. Arora and Arora[8] emphasize that fully anonymized synthetic patient data can "replace the use of real patient data in certain contexts." In our work, all

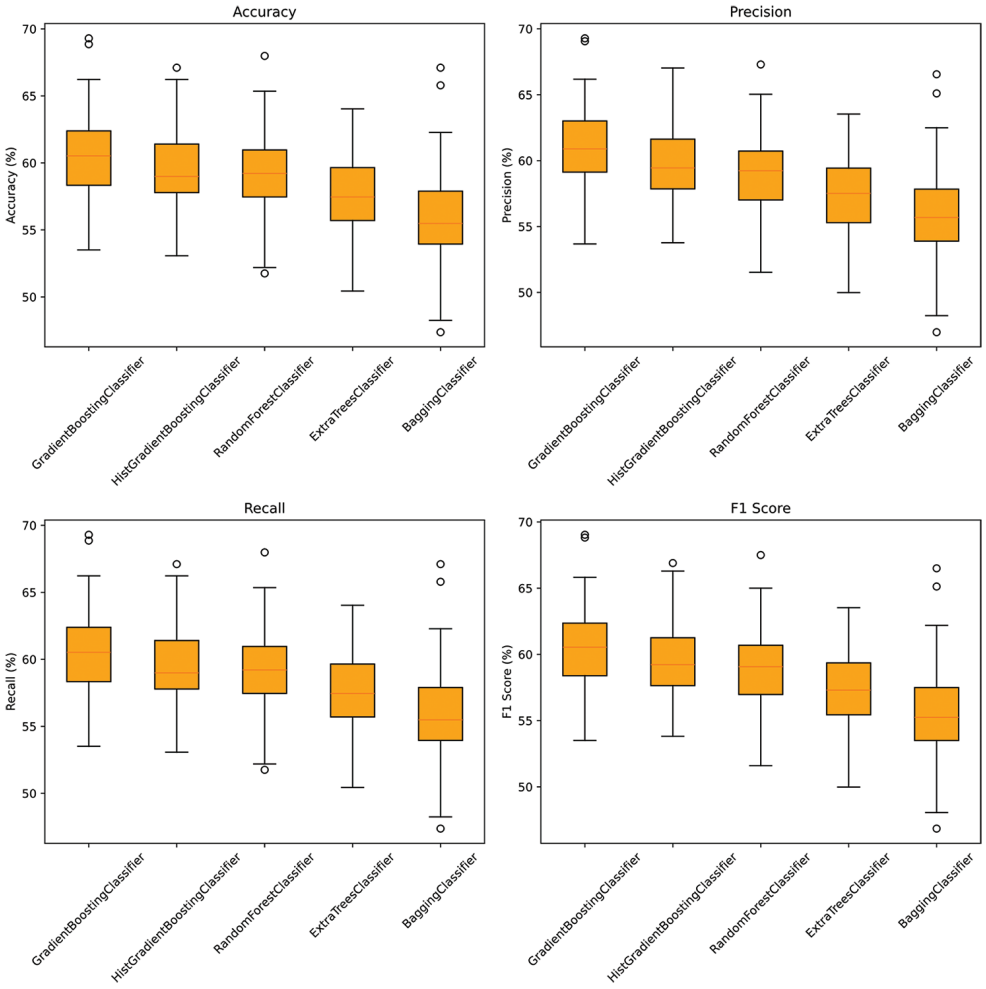

Figure 30. Performance metrics plots of the five most successful classifiers on the conditional tabular generative adversarial network dataset (excluding height and weight attributes)

synthetic examples were derived from the real EOL data, but in principle, such models could be used to generate new plausible patient profiles. This could allow researchers to share or analyze tabular health data while preserving privacy or to simulate large cohorts for training more complex models.

The EOL dataset is cross-sectional and self-reported, which limits causal inference. The synthetic data quality was not evaluated beyond model performance; future work could apply standardized metrics to quantitatively assess the resemblance and privacy of generated samples. We also noted that CTGAN's underperformance may be due to the limited data size; experimenting with larger or multi-source datasets could test whether GANs become more reliable under such conditions. Clinically, while our accuracy without BMI (~75% F1) is promising, it may not be sufficient for a standalone diagnosis. Rather, it suggests that such models could serve as preliminary screening tools to flag at-risk individuals for further evaluation.

In summary, our study demonstrates that ML classifiers for obesity can be trained effectively on augmented synthetic data, even when key anthropometric features are absent. This has practical relevance for nutritional and clinical practice, as it implies that an AI tool could estimate obesity risk from just diet and lifestyle information (e.g., survey responses) with reasonable accuracy. It also highlights that synthetic data generation is a viable strategy to mitigate data limitations in health research.

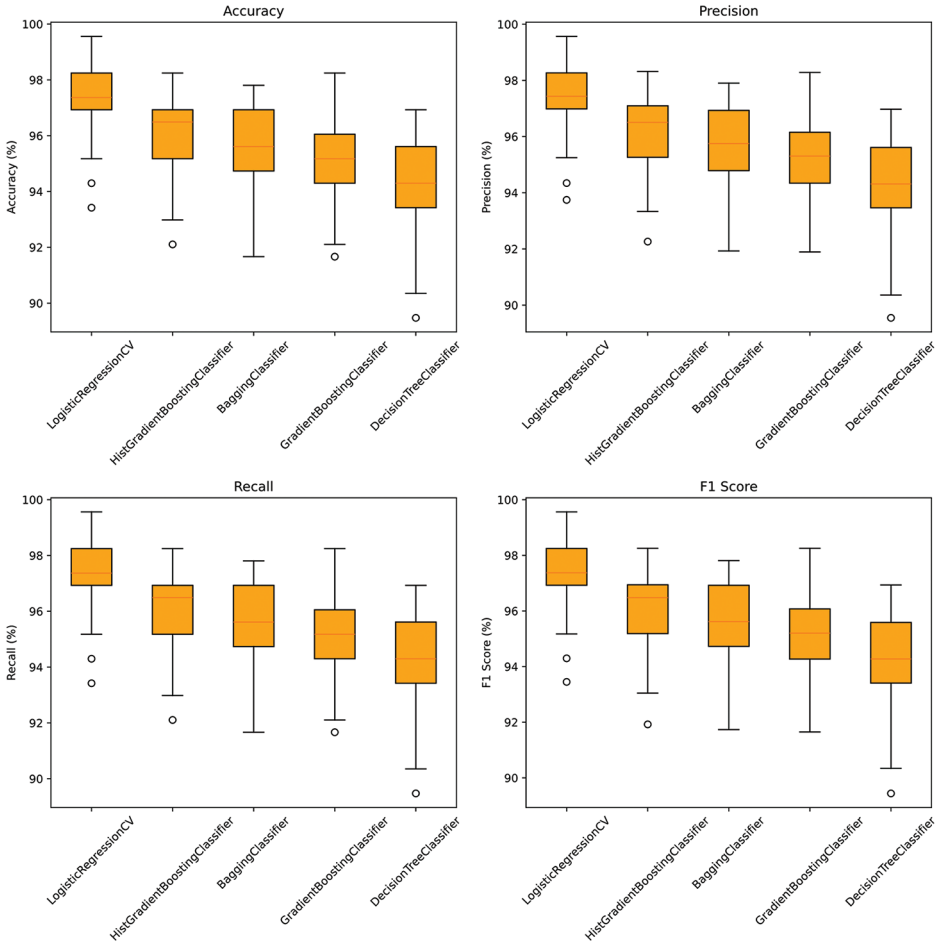

Figure 31. Plots of the performance metrics of the five most successful classifiers on the conditional tabular generative adversarial network dataset (using height and weight attributes)

5. Conclusion and future work

This study demonstrates the effectiveness of training classification models using synthetic data generated through techniques such as SMOTE-NC and TVAE, even when the original dataset is limited in size. A detailed analysis revealed that favorable classification performance can be achieved without the inclusion of height and weight attributes when using synthetic datasets generated by SMOTE-NC and TVAE. However, for the dataset generated using CTGAN, excluding height and weight features results in suboptimal model performance. In contrast, incorporating these features yields significantly improved results across all three datasets, with F1-scores approaching 100%. These findings are particularly important for obesity level prediction, as they indicate that even in the absence of direct anthropometric measures such as height and weight, synthetic data generated using appropriate techniques can support the development of reasonably accurate models – especially with SMOTE-NC and TVAE. While SMOTE remains a widely adopted technique in the literature for synthetic data generation, this study also highlights the viability of NN-based approaches such as TVAE. In particular, classifiers trained on SMOTE-NC and TVAE datasets (excluding height and weight) achieved an F1 score of approximately 75% on the test set – an outcome not replicated with CTGAN-generated data. Future research directions include: (i) Exploring CTGAN and other generative models on larger or more diverse obesity datasets to improve synthetic fidelity; (ii) integrating additional predictive features (e.g., genetic, microbiome, or detailed metabolic biomarkers) to enhance model relevance; and (iii) conducting prospective validation of synthetic-data-augmented models in clinical or community cohorts to assess their real-world utility in preventive health. We believe that the continued development of synthetic tabular data methods will strengthen AI-driven obesity prevention and nutrition research.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: All authors
Formal analysis: All authors
Investigation: All authors
Methodology: Hakan Alp Eren, Sinem Bozkurt Keser
Writing – original draft: All authors
Writing – review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The dataset used in this study is publicly available from the University of California, Irvine ML Repository under the title *Estimation of Obesity Levels Based on Eating Habits and Physical Condition*: https://archive.ics.uci.edu/dataset/544.

References

1. Ural D, Kılıçkap M, Göksülük H, *et al*. Data on prevalence of obesity and waist circumference in Turkey: Systematic review, meta-analysis and meta-regression of epidemiological studies on cardiovascular risk factors. *Turk J Cardiol Arch*. 2018;46(7):577-590.

   doi: 10.5543/tkda.2018.62200

2. Yavuz R, Tontuş H. The clinical approach to the obesity in adult, adolescent and pediatric age groups. *J Exp Clin Med*. 2013;30(1s):69-74.

3. Rosengren A. Obesity and cardiovascular health: The size of the problem. *Eur Heart J*. 2021;42(34):3404-3406.

   doi: 10.1093/eurheartj/ehab518

4. Dönder E, Önalan E. Definition, epidemiology, and clinical evaluation of obesity. *Fırat Med J*. 2018;23(3):1-4.

5. UCI Machine Learning Repository. *Estimation of Obesity Levels Based On Eating Habits and Physical Condition*. Irvine: UCI Machine Learning Repository; 2019.

   doi: 10.24432/C5H31Z

6. Shi R, Wang Y, Du M, Shen X, Wang X. *A Comprehensive Survey of Synthetic Tabular Data Generation*. [arXiv Preprint]; 2025.

   doi: 10.48550/arXiv.2504.16506

7. Hernadez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions. *Methods Inf Med*. 2023;62(S01):e19-e38.

   doi: 10.1055/s-0042-1760247

8. Arora A, Arora A. Generative adversarial networks and synthetic patient data: Current challenges and future perspectives. *Fut Healthc J*. 2022;9(2):190-193.

   doi: 10.7861/fhj.2022-0013

9. Sámano R, Lopezmalo-Casares S, Martínez-Rojano H, *et al*. Early life determinants of overweight and obesity in a sample of Mexico city preschoolers. *Nutrients*. 2025;17(4):697.

   doi: 10.3390/nu17040697

10. Sobas K, Suliga E, Bryk P, Gluszek S. Dietary patterns and nutritional status in bariatric surgery candidates-a cross-sectional study. *Nutrients*. 2025;17(4):716.

    doi: 10.3390/nu17040716

11. Colonnello E, Libotte F, Masi D, *et al*. Eating behavior patterns, metabolic parameters and circulating oxytocin levels in patients with obesity: An exploratory study. *Eating Weight Disord*. 2025;30(1):6.

    doi: 10.1007/s40519-024-01698-w

12. El-Sehrawy AAMA, Khachatryan LG, Kubaev A, *et al*. Triglyceride-glucose index: A potent predictor of metabolic risk factors and eating behavior patterns among obese individuals. *BMC Endocr Disord*. 2025;25(1):71.

    doi: 10.1186/s12902-025-01887-3

13. Kuckuck S, Van der Valk ES, Lengton R, *et al*. Long-term hair cortisone and perceived stress are associated with long-term hedonic eating tendencies in patients with obesity. *Psychoneuroendocrinology*. 2025;171:107224.

    doi: 10.1016/j.psyneuen.2024.107224

14. Palechor FM, De la Hoz Manotas A. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data Brief*. 2025;25:104344.

    doi: 10.1016/j.dib.2019.104344

15. Helforoush Z, Sayyad H. Prediction and classification of obesity risk based on a hybrid metaheuristic machine learning approach. *Front Big Data*. 2024;7:1469981.

    doi: 10.3389/fdata.2024.1469981

16. Ayub H, Khan MA, Shehryar Ali Naqvi S, *et al*. Unraveling the potential of attentive Bi-LSTM for accurate obesity prognosis: Advancing public health towards sustainable

cities. *Bioengineering (Basel)*. 2024;11(6):533.

doi: 10.3390/bioengineering11060533

17. Shakti MAS, Vijayalakshmi M, Kumar N, Vaidhehi M. Analysis on Various Machine Learning Framework for Obesity Level Prediction. In: *Proceedings of the 1st International Conference on Contemporary Global Challenges and Urban Innovations (ICCGUI) IEEE*. Vol. 1; 2024. p. 406-411.

doi: 10.1109/IC-CGU58078.2024.10530812

18. Yağmur N. A hybrid approach to obesity level determination with decision tree and pelican optimization algorithm. *J Sci Rep A*. 2024;57:97-109.

doi: 10.59313/jsr-a.1447814

19. Özkurt C. Examination and evaluation of obesity risk factors with explainable artificial intelligence. *Comput Electron Med*. 2024;1(1):12-17.

doi: 10.69882/adba.cem.2024072

20. Wang X. Predicting obesity risk through lifestyle habits: A comparative analysis of machine learning models. *E3S Web Conf*. 2024;385:05037.

doi: 10.1051/e3sconf/202455305037

21. Okpe OA, Odey JA, Abiodum OJ. A novel multi-class classification of obesity level using artificial neural network. *Int J Adv Multidiscip Res Studies*. 2024;4(3):1374-1379.

22. Azad M, Khan MFK, El-Ghany SA. XAI-enhanced machine learning for obesity risk classification: A stacking approach with LIME explanations. *IEEE Access*. 2025;13:13847-13865.

doi: 10.1109/ACCESS.2025.3530840

23. Solomon DD, Khan S, Garg S, *et al*. Hybrid majority voting: Prediction and classification model for obesity. *Diagnostics* (*Basel*). 2023;13(15):2610.

doi: 10.3390/diagnostics13152610

24. Kaur R, Kumar R, Gupta M. Predicting risk of obesity and meal planning to reduce obesity in adulthood using artificial intelligence. *Endocrine*. 2022;78(3):458-469.

doi: 10.1007/s12020-022-03215-4

25. Muliawan A, Fauziah DA, Afrianto E. Obesity risk prediction using random forest based on eating habit parameters. *INSIDE J*. 2024;2(1):13-18.

26. Choudhuri A. A Hybrid Machine Learning Model for Estimation of Obesity Levels. In: *Proceedings of the International Conference on Data Management, Analytics and Innovation*. Vol. 137; 2023. p. 414-423.

doi: 10.1007/978-981-19-2600-6_22

27. Cervantes RC, Palacio ALH. Estimation of obesity levels based on computational intelligence. *Inf Med Unlocked*. 2020;21:100472.

doi: 10.1016/j.imu.2020.100472

28. Ganie SM, Reddy BB, Rege M. An investigation of ensemble learning techniques for obesity risk prediction using lifestyle data. *Decis Analyt J*. 2025;14:100539.

doi: 10.1016/j.dajour.2024.100539

29. Nagarajan SG, Balasubramanian V, Gonugunta P, Gudla SK. Obesity level prediction using deep learning approach-a comparative analysis. *Eng Appl Sci Res*. 2024;51(4):540-554.

30. Umoh PN, Nneji GU, Monday HN, *et al*. Optimizing machine learning classifiers and feature selection techniques for obesity levels estimation using physical habits and dietary data. *World Sci News*. 2024;198:326-353.

doi: 10.1142/WSN198(2024)325-353

31. Vairachilai S, Periyanayagi S, Raja SPR. PIPR machine learning model: Obesity impact analysis. *Open Biomed Eng J*. 2024;18(1):1-20.

doi: 10.2174/0118741207289421240430115207

32. Forte P, Encarnação S, Monteiro AM, *et al*. A deep learning neural network to classify obesity risk in portuguese adolescents based on physical fitness levels and body mass index percentiles: Insights for national health policies. *Behav Sci*. 2023;13(7):522.

doi: 10.3390/bs13070522

33. Yağın FH, Gülü M, Görmez Y, *et al*. Estimation of obesity levels with a trained neural network approach optimized by the Bayesian technique. *Appl Sci*. 2023;13(6):3875.

doi: 10.3390/app13063875

34. Gözükara Bağ HG, Yağın FH, Görmez Y, *et al*. Estimation of obesity levels through the proposed predictive approach based on physical activity and nutritional habits. *Diagnostics*. 2023;13(18):2949.

doi: 10.3390/diagnostics13182949

35. Yang Y, Khorshidi HA, Aickelin U. A review on over-sampling techniques in classification of multi-class imbalanced datasets: Insights for medical problems. *Front Digit Health*. 2024;6:1430245.

doi: 10.3389/fdgth.2024.1430245

36. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res*. 2017;18(17):1-5.

37. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular Data using Conditional GAN. In: *Advances in Neural Information Processing Systems*; 2019. p. 32. Available from: https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-abstract.html [Last accessed on 2024 Dec 12].

38. Patki N, Wedge R, Veeramachaneni K. The Synthetic Data Vault. In: *International Conference on Data Science and Advanced Analytics (DSAA)*; 2016. p. 399-410.

doi: 10.1109/DSAA.2016.49

39. Luo Y, Tao J, Zhu Y, Xu Y. HSS: Enhancing IoT malicious traffic classification leveraging hybrid sampling strategy. *Cybersecurity*. 2024;7(1):11.

doi: 10.1186/s42400-023-00201-9

40. Yadav P, Gaur M, Madhukar RK, Verma G, Kumar P. Rigorous experimental analysis of tabular data generated using TVAE and CTGAN. *Int J Adv Comput Sci Appl*. 2024;15(4):1250-1262.

doi: 10.14569/ijacsa.2024.01504125

41. Huang GL, Wu PY. CTGAN: Cloud transformer generative adversarial network. In: *2022 IEEE International Conference on Image Processing (ICIP);* 2022. p. 511-515.

doi: 10.1109/ICIP46576.2022.9897229