Engineering and Applied Science Research 2024;51(4):540-554 Research Article

EASR

Engineering and Applied Science Research

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

Obesity level prediction using deep learning approach – A comparative analysis

Srinivasa Gupta Nagarajan[1)], Valarmathi Balasubramanian*[2)], Phani Gonugunta[2)] and Saran Kumar Gudla[2)]

[1)]Department of Manufacturing Engineering, School of Mechanical Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India
[2)]Department of Software and Systems Engineering, School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, Tamil Nadu, India

Received 4 January 2024
Revised 2 July 2024
Accepted 5 July 2024

Abstract

Obesity, the excessive accumulation of body fat, affected millions globally and was influenced by eating habits, lack of acti vity, genetics, environmental factors, and emotional strain. It could lead to severe health issues, including insulin resistance, cardiovascular diseases, cancer, sleep apnea, joint problems, and mental health disorders. This study aimed to predict obesity levels using Machine Learning (ML) and Deep Learning (DL) models on a real-life dataset of obesity patients. The dataset comprised several patient health records with 17 different elements related to obesity, classifying obesity levels into seven types. The study evaluated the accuracy of various models before and after applying the Synthetic Minority Over-sampling Technique (SMOTE). Before SMOTE, the TabNet (T) and XG-Boost (XGB) classifiers achieved high accuracies of 96.6% and 96.2%, respectively, outperforming Random Forest (RF) (94.8%), Multi-Layer Perceptron (MLP) (94.5%), Bagging (B) (94.07%), Decision Tree (DT) (93.6%), Support Vector Machine (SVM) (82.5%), K-Nearest Neighbor (KNN) (75.9%), Stochastic Gradient Descent (SGD) (68.6%), AdaBoost (AB) (28.4%), Stacking (S) (16.8%), and G-Boost (GB) (95.5%). After applying SMOTE, GB and XGB showed improved accuracies of 99.3% and 99%, respectively, surpassing RF (97.4%), Bagging (96.28%), DT (96.9%), SVM (90.3%), KNN (85.7%), SGD (67.6%), AB (34.9%), and Stacking (12.3%). Comparatively, the existing methods showed accuracies with GB (97.2%), DT (96.7%), RF (94.8%), SVM (43.4%), and AB (33.1%), while the proposed models exhibited superior performance: GB (99.3%), DT (96.9%), RF (97.4%), SVM (90.2%), AB (34.9%), XGB (99%), TabNet (98.4%), and MLP (97.7%). The proposed models significantly outperformed the existing ones, demonstrating their effectiveness in predicting obesity levels.

Keywords: Obesity level, SMOTE, TabNet Classifier, XGB Classifier, GB Classifier

1. Introduction

Obesity defined as an unusual degree of fatness and a condition that is diagnosed by a BMI of 30 or higher, is a developing concern in the globe. Currently, more than 650 million grownups worldwide are reported to be obese and the World Health Organization (WHO) predicts that globally, more than one billion individuals may be obese by the year 2030. Similar to adult populations, childhood overweight, and obesity has increased; as per 2016 statistics, over 340 million children and adolescents below the age of 19 are overweight or obese, while in early ages particularly in the age group of 0-5 there has been a significant change in obesity statistics; in 1990 there were 32 million obesity cases but this figure rose to 41 million in 2016.

It is known that obesity is influenced by a number of factors which reflect diet, lifestyle, genetic makeup, environment, and psychology. Dietary behaviors include taking in foods that are high in energy, low in nutrients and typically processed and containing high-sugar beverages; these foods are available and promoted. The lifestyles are founded on habits and these include a lack of exercise due to the civilized and modern ways of life that do not require a lot of physical work. There are certain inheritance patterns that run in families and they determine how a person’s body handles fat. For instance, poor diet and lack of physical exercise because of environmental endowment that as unhealthy food and lack of playgrounds or other exercising areas coupled with poor income status for those in the urban and low-income bracket. Furthermore, matters concerning psychology like stress, depression, and particular eating disorders significantly contribute to the causation and exacerbation of obesity as they most of the time prompt emotional eating and poor food selection.

Obesity can be described as a major risk factor as it leads to various non-communicable diseases in patients. These include cardiovascular diseases inclusive of heart disease and stroke, type 2 diabetes, some cancers like breast, colon, and endometrial cancer, and musculoskeletal diseases like osteoarthritis. It has also been rated to have several other related factors involving breathing difficulties such as sleep apnea, and liver diseases and it even results in a reduced quality of life, and even the death rate is known to be high among obese persons. It is combated using several strategies that include improving proper diet, encouraging timely exercises, fostering an enabling environment, and not forgetting the psychological components.

*Corresponding author.
Email address: valargovindan@gmail.com
doi: 10.14456/easr.2024.51

Machine learning (ML) can be defined as a subdiscipline of artificial intelligence that allows systems to learn from data and make predictions or decisions, which results in substantial advantages throughout various industries due to the possibility of discovering hard-to-find dependencies in large sets of data. Gathering and preprocessing the data, there is also feature extraction and finally, there is a division of data into training data, testing data, and validation data. It is critical for the best model selections concerning the nature of problems, whether regression, classification, clustering, or deep learning.

XGB is a variant of GBM that is well-suited when working with rather structured data and tends to be one of the most accurate and robust. They establish decision trees in sequences in an attempt to refine mistakes step by step; the likelihood is regulated by invoking techniques that protect from overemphasizing certain features, making their model rich in efficiency and flexibility.

This is the process of aggressive learning to uses a lot of weak learners to come up with a single strong learner that can perform a classification or regression task. It is a process of iteratively introducing a weak learner, training on the residual, and continuing this process until a termination condition is fulfilled, thus creating a strong model that may predict very well.

As closely related methods, XGB is highly acclaimed for its precision and ability to solve a range of intricate issues, as well as gradient boosting.

The major contributions of the proposed works are:

1. Ten machine learning model classifiers and two deep learning model classifiers are used in this study, and several techniques like as Oversampling and Hyperparameter tweaking are utilised to achieve optimal accuracy.
2. The accuracy of our study is 99.3% greater than the accuracy of the current model on the same particular subject, which was conducted by other researchers.
3. A cloud-hosted webpage built using this study effort computes the obesity level based on the real-time data provided as input.
4. Among the twelve classifiers, GB, XGB, TabNet, MLP, and RF show the maximum accuracy of 99.3%, 99%, 98.4%, 97.7%, and 97.4% respectively.

The paper lays down a step-by-step framework for using machine learning, importing necessary packages, and loading the data in an integrated development environment. It pays special attention to data pre-processing to have balanced datasets and enough data to work with and after that entails an exhaustive analysis of the data and validation of results. Data visualization helps to explore the characteristics of the given dataset, and scaling improves the model's performance. The given data set can be divided into training and testing data sets which help in model building using various classification techniques. One of the essential modifications is the use of the SMOTE algorithm to help balance the datasets with the help of statistical tables and graphs.

The paper is structured into six sections: The research background is defined in the Introduction and the Literature Survey (Section 1) contains a critical analysis of the prior studies, Dataset Description (Section 2) describes the dataset explaining important characteristics with the help of figures, and tables. Methodology (Section 3) covers the algorithms used in the framework of the study and presents the proposed solution in the form of a flowchart. Results (Section 4) present the graphs and tables to show the differences in the accuracies between the model before and after the implementation of SMOTE. In the final section called Conclusion and Future Scope (Section 5), the earlier analysis is brought together, backward and forward explorations are made, recommendations are given, and directions for future studies are highlighted. About the whole text, it is necessary to notice that all the sources used in the work are listed in References (Section 6), which guarantees the scholarly approach and is informational for further research. Thus, aims are rationalized to uncover how SMOTE affects the accuracy of a model and the balance of a dataset to aid future research in machine learning and data science.

This work compares the effectiveness of a dataset before and post SMOTE algorithm use. The use of the SMOTE algorithm is the primary innovation; in contrast to other publications, the SMOTE algorithm has been utilized to balance a dataset after research on the effects of imbalances in datasets. The whole process has been carefully documented with stats and graph visualizations that illustrate each method, and each outcome parameter, along with each dataset - in forms that are both balanced and imbalanced.

The description of various forms of obesity with detailed codes is possible according to the ICD, which means that diagnosis and classification are possible. Here are the definitions for obesity in both ICD-10 and ICD-11:

ICD-10 has 5 subtypes related to the obesity (E66) are E66.0, E66.1, E66.2, E66.3, E66.8, and E66.9

E66.0: Obesity due to excess calories - Obesity is mainly due to the difference between the total amount of consumed energy and the amount of energy used up by the human body. This category includes obesity resulting from overeating and also inhaling foods without performing many activities.

E66.1: Drug-induced obesity - Various diseases that are related to obesity as a result of taking some medicines that cause weight gain. This can include drugs such as particular antianxiety, antidepressants, antipsychotics, and corticosteroid drugs.

E66.2: Extreme obesity with alveolar hypoventilation - Non-aversive somatometric obesity is also referred to as obesity hypoventilation syndrome which is characterized by extreme obesity in conjunction with decreased breathing and high levels of carbon dioxide. This condition should be well-regulated because it causes health risks in the respiratory and cardiovascular systems.

E66.3: Overweight - It is a state where an individual is overweight but not obese; hence the mass index is slightly higher than the normal average. Being overweight destroys the natural balance of your body and increases the risk of obesity-related diseases.

E66.8: Other obesity - This classification also has other particular types of obesity that are not included in other categories. This category includes forms that are not so frequent like obesity related to genetic diseases.

E66.9: Obesity, unspecified - An umbrella category for obesity in which the root of the condition is not known. This code is used when detailed information regarding the type or the cause of obesity cannot be ascertained.

ICD-11 has 6 subtypes related to the Obesity (5B81) are 5B81.0, 5B81.1, 5B81.2, 5B81.3, 5B81.Y, and 5B81.Z.

5B81.0: Obesity as a result of more calories - In particular, obesity is a disease connected with the disturbance of the relationship between energy consumption and energy utilization. This kind of obesity is probably, in many cases, attributed to diet and physical activity.

5B81.1: Drug-induced obesity - Obesity as a side effect of medications: This is where some medications tend to have negative impacts as far as weight is concerned. Examples of drugs related to weight gain are antipsychotics, antidepressants, and antiepileptics.

5B81.2: Extreme obesity with alveolar hypoventilation - Formerly referred to as Pickwickian, the disease is an obesity hypoventilation syndrome characterized by severe obesity and breathing problems or high levels of carbon dioxide in the blood. The treatment measures are often focused on weight reduction processes and respiratory care.

5B81.3: Overweight - A state of having an excess size of body mass that is above the ideal weight for a specific height but below the level of obesity. Obesity could also be prevented in overweight people through modification of their lifestyles.

5B81.Y: Other specified obesity - Other specific types of obesity that cannot be described as the first or second degree. This is with respect to special types of obesity or obesity-related to certain diseases.

5B81.Z: Obesity (unspecified) - A broad diagnosis for obesity particularly when the precise nature or exact cause has not been described. This code is used where information about the type or cause of obesity as defined as detailed above cannot be ascertained or is irrelevant.

Obesity is considered a global epidemic which makes an effect on all populations [1]. Manual record analysis of the prognosis is highly inaccurate and will lead to disease progression if early diagnosis is not done [2, 3]. ML approaches namely RF, KNN, SVM, GB, DT, Logistic regression, and deep learning models have been used for the purpose of identifying obesity status and risk factors [2-10].

Maria et al. [2] attained a 99 percent accuracy in their experiment. 5% accuracy with the GB classifier their performance was 95. 74% with KNN. Alsareii et al. [4] applied AI together with EHR to diagnose the condition of obesity, with a level of accuracy of 98. 5% and 99. 6% success rates. The following barriers to weight management had been proposed by scholars; Yagin et al. [5]. The factors include; Activity level and dietary patterns. Jeon et al. [11] obtained the KNHANES survey data to estimate the obesity risk using triglycerides, ALT, glycated haemoglobin, and uric acid.

Celik et al. [8] obtained 97.8 percent classification accuracy with the help of the cubic SVM algorithm. Cui et al. [9] classified people into weight categories with Decision Tree, Logistic Regression, and KNN. Molina et al. [12] have established a model for estimating body fat, and the results yielded 96. As mentioned earlier, women in INTER Chinese reported an overall 65% accuracy rate with LMT. Rodriguez et al. [13] employed RF to accurately detect overweight/obese people.

Pinto et al. [14] also identified the findings of the study DT 71. 68% accurate for physical activity data according to the researchers who developed the application; another study identified it to be 54% accurate for dietary data and 63%. 63% for fitness data. Cervantes and Palacio [15] described factors concerning obesity by means of DT, SVM, and Simple K-means techniques. Deniz et al. [16] used the SEMMA techniques assisted with J48 DT, which has yielded rather high results. De-La-Hoz-Correa et al. [17] tested childhood obesity using the SEMMA technique acquiring high accuracy in J48 DT. These authors [18] presented obesity rates for Mexico, Peru, and Colombia that would allow for the detection tools.

DeGregory et al. [19] compared the performance of DT, logistic regression, and neural networks and used national health surveys. Montañez et al. [20] employed the genetics of ML systems to predict people’s BMI status with a 90 percent accuracy. 5% AUC with SVM. Thus, according to the research of Yahia et al. [21], gender-oriented programs for health promotion were based on BMI and dietary behavior. Some researchers/works Dugan et al. [22] predicted childhood obesity based on the RF, Random Tree, J48, ID3, and Naïve Bayes classifiers. Davila-Payan et al. [23] adopted a regression equation to forecast BMI from national polls and census information. Muhamad Adnan et al. [24] introduced the Naive Bayes with Genetic Algorithm composite model for childhood obesity with increased accuracy by up to 75%. About the use of data mining tools to predict childhood obesity Muhamad Adnan et al. [25]

Some aspects, for instance, bilirubin were still worth providing a closer look at concerning the classification of obesity [26]. Recent studies pointed out that there is a necessity for more sophisticated models that allow for the physiological, psychological, and sociocultural characteristics [27]. Nicklas et al. [28] tried to consider eating bad habits and obesity in children to direct certain measures for the meal patterns of consumption. Zhao et al. [29] suggested in this work and they looked at the relationship between eating habits and weight issues as hypertension in a group of 34,040 individuals aged 45 and above who participated in the Nutrition and Health Surveillance in China from 2015 to 2017.

In their study, Alkhalaf et al. [30] suggested that other such approaches be further investigated in subsequent research, such as neural networks and deep learning. In their work, Devi et al. [31] stated the utilization of data from institutions in Tamil Nadu and the UCI Machine Learning Repository. The findings indicated that Logistic Regression achieves prediction accuracy in Decision Tree and other models. Maria et al. [2] proposed that they meant to enhance public clinical datasets, which indicates there is room for advancement in the way data is gathered and employed and to work on an imbalanced dataset. Yagin et al. [5] introduced hyperparameters and used the Bayesian optimization method to adjust them and therefore it could be argued that there was room for improving the model accuracy again through hyperparameters tuning. Muhamad Adnan et al.'s [24] study used Naive Bayes and Genetic Algorithm to predict childhood obesity, showing a 75% improvement in accuracy in their initial experiment, suggesting potential for further improvement with advanced techniques. Pinto et al. [14] - Reported accuracy rates of 71.54% and 63.63% for Decision Trees on dietary and fitness data respectively, and 65.85% and 69.32% for K-Nearest Neighbors. These accuracies, being in the 60-70% range, could potentially be improved with more advanced techniques. Rodriguez et al. [13], Their Random Forest model yielded accuracy, precision, recall, and f1-score at 78%. While this was a decent accuracy, it was not extremely high, suggesting there might be room for improvement. Dugan et al. [22], while they didn't provide specific accuracy figures, they mentioned examining six different ML techniques to construct a sensitive model, which might imply they were looking for ways to improve performance. Palechor and de la Hoz Manotas [18] stated that "The researchers leveraged the Weka application and the SMOTE filter to generate 77% of the data, with the remaining 23% obtained directly from people via a web platform." and suggested the importance of the smote technique and also suggested that the smote technique should be used properly to see best results mentioning that it is an imbalanced dataset. DeGregory et al. [19] investigated various machine learning methods, including artificial neural networks and deep learning techniques and the inclusion of these advanced techniques implies their potential for improving obesity prediction and classification. Kivrak [1] worked on applied deep-learning algorithms to predict levels of obesity. And suggested the use of deep learning implies interest and advancement in advanced techniques for obesity prediction.

Authors revealed the following limitations in the prior research works: low accuracy [5, 13, 14, 22, 24], imbalanced data [2, 18], and the lack of deep learning and Neural network algorithms [1, 19, 30]. To overcome these issues, the proposed work has effectively incorporated the SMOTE (Synthetic Minority Over-sampling Technique) algorithm to balance up the data, which has helped in boosting the performance of the model substantially. Furthermore, by using new deep learning methods included in development, the proposed work succeeded in eliminating the above-mentioned shortcomings while making the achieved performance indicators considerably higher. These methods were applied in this work not only to fix the problems of imbalanced data but also to utilize the potential of deep learning to provide reliable and accurate outcomes.

2. Dataset description and sample data

URL to the dataset this study was done with: https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

However, due to their eating habits and physiological circumstances, this dataset assesses the prevalence of obesity among Mexicans, Peruvians, and Colombians. Overall, 2111 separate records may be created depending on 17 different attributes. Every record is labeled with the variable class NObesity (Obesen _Level), which allows the data to be classified using the following values: Overweight_Level-I (over_weight-I), Overweight_Level-II (over_weight-II), obesity_Level-I (obesity-I), Insufficient_Weight (Insufficient_Weight), Normal_weight (normal), obesity_Level-III (obesity-III), obesity_Level-II (obesity-II).

After calculating each individual's mass body index, the findings were compared to WHO and Mexican norms [18].

$$Body\ Mass\ Index\ (BMI) = \frac{Weight}{Height*Height} \tag{1}$$

The class labels are defined based on BMI with the corresponding values given below.

- insufficient_weight < 18.5
- normal is from 18.5 - 24.9
- over_weight-I is from 25.0 - 27.4
- over_weight-II is from 27.5 - 29.9
- obesity-I is from 30.0 - 34.9
- obesity-II is from 35.0 - 39.9
- obesity-III > 40

Two-thirds of the total data were sourced from consumers via online platforms with the last quarter having been generated artificially with the help of the Weka tool and SMOTE algorithm.

- Number of Attributes = 17 (16 Independent & 1 Dependent)
- Total Number of Records = 2111 records
- Number of classes in Dependent attribute column = 7 classes

The independent variables are Frequent_intake_of_high-calorie_foods (FAVC), Drinking_Alcohol (CALC), Frequency_of_Vegetable_Consumption (FCVC), frequency_of_main_meals (NCP), Consumption_of_Food_in_Between_Meals (CAEC), Water_intake_on_a_daily_basis (CH20), tracking_of_caloric_intake (SCC), Frequency_of_physical_actions (FAF), Duration_of_electronic_gadget_use (TUE), Type_of_Transportation (MTRANS), Gender, Age, Height, Weight, Family history with overweight, and Smoke.

The original dataset's composition is unbalanced. A balanced dataset is created by the SMOTE algorithm application to imbalanced data. This method makes use of oversampling. In order to verify that each class label has an equal number of rows, or more or fewer rows, in the dataset, any classes with insufficient rows are supplied by extra records. A dataset that is unbalanced has asymmetry. A skewed class distribution from an unbalanced dataset has several effects on the accuracy of the models.

Thus, it is imperative to balance the facts. By oversampling the positive class label, the accuracy of the findings can be increased. The SMOTE algorithm is used for oversampling in this study. By establishing its model on nearest neighbors, the SMOTE algorithm increases the occurrence of the minority class or minority class group in the given dataset.

Table 1 shows the size of the dataset before and after applying SMOTE algorithm each class or each type of obese level is either having 499 records or 500 records which combines to 3496 Records in total, whereas before applying SMOTE algorithm, the total records used to be 2111 with a varying number of records for each obesity level.

Table 1 Size of dataset Prior to and post SMOTE algorithm Comparison

Obese_level	Imbalanced dataset size (Before applying SMOTE algorithm)	Balanced dataset size (After applying SMOTE algorithm)
1 normal	287	499
2 over_weight-I	290	500
3 over_weight-II	290	500
4 obesity-I	351	499
5 Insufficient_Weight	272	499
6 obesity-II	297	500
7 obesity-III	326	499
Sum	2111	3496

Some of the dataset attributes relationship with NObeyesdad are mentioned below. Figure 1 represents the relationship between Age and NObeyesdad attributes. Age attribute values lie between 10 to 65. NObeyesdad attribute values are Insufficient_Weight, normal, over_weight-I, over_weight-II, obesity-I, obesity-II, and obesity-III. Figure 2 represents the relationship between Age and NObeyesdad attributes. Weight attribute values lie between 39 to 173. NObeyesdad attribute values are Insufficient_Weight, normal, over_weight-I, over_weight-II, obesity-I, obesity-II, and obesity-III.

Figure 3 represents the relationship between CALC and NObeyesdad attributes. CALC attribute values are no, sometimes, frequently, and always. NObeyesdad attribute values are Insufficient_Weight, normal, over_weight-I, over_weight-II, obesity-I, obesity-II, and obesity-III. Figure 4 represents the relationship between MTRANS and NObeyesdad attributes. MTRANS attribute values are Public_transportation, walking, automobile, motorbike, and bike. NObeyesdad attribute values are Insufficient_Weight, normal, over_weight-I, over_weight-II, obesity-I, obesity-II, and obesity-III.

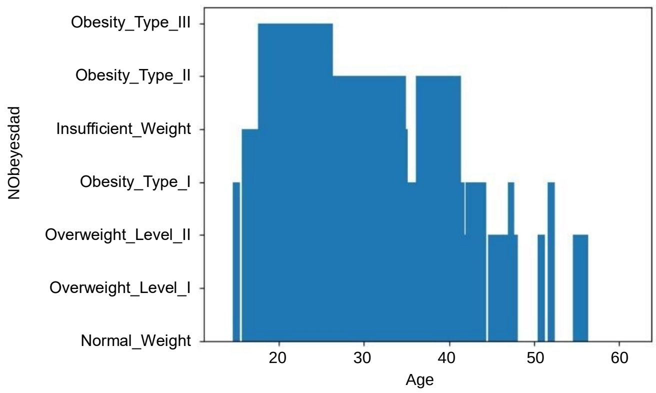

Figure 1 Relationship between Age and NObeyesdad

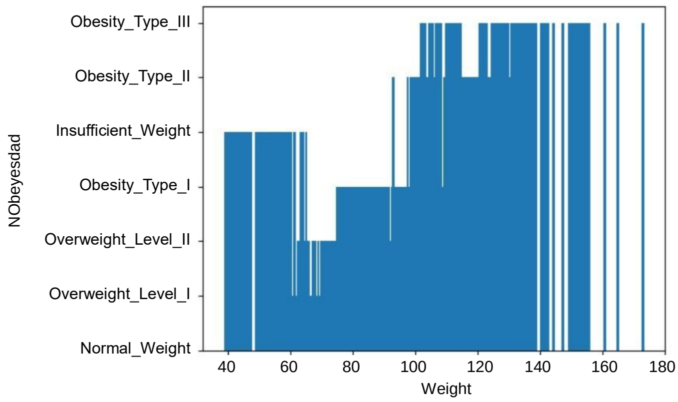

Figure 2 Relationship between Weight and NObeyesdad

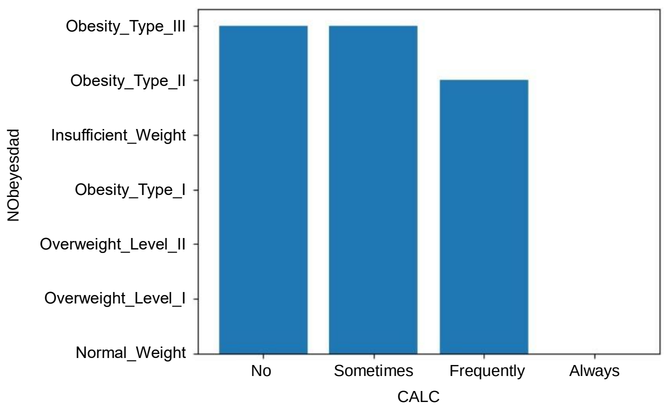

Figure 3 Relationship between CALC and NObeyesdad

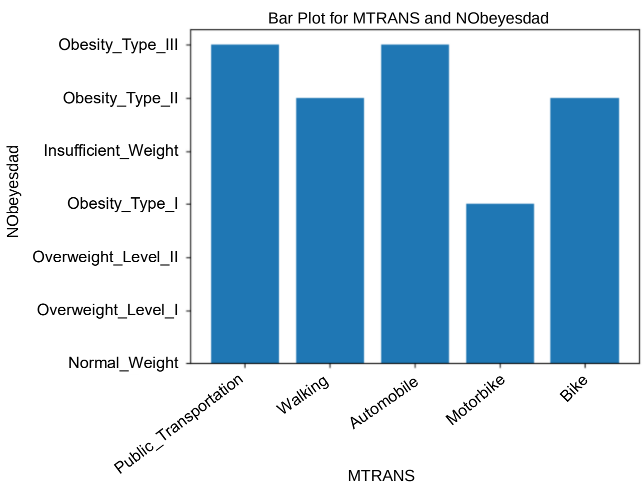

Figure 4 Relationship between MTRANS and NObeyesdad

3. Methodology

This study proposes and applies some common ML techniques to those algorithms able to measure obesity levels in an accurate manner. The prediction is made upon simple info like gender, age, height, weight, a genetic record of overweightness and food behaviors like taking high-energy products, fruits, principal courses, snacks in between meals, and constant water supplies. There are several physical state variables such as CALC and smoking, while the others are numerical indicators like eating habits, physical exercises, and transportation modes that contribute to the body condition (physical state). Ten different categorization Machine algorithms are used by the provided approaches to compare the accuracy values in this paper, including KNN, SVM, DT, RF, GB, AB, SGD, XGB, Bagging, and Stacking to estimate the level of obesity. As a result, the ten algorithms in our proposed study will be evaluated to determine which are the most accurate and efficient**.** Also, the entire dataset was divided into an 80:20 ratio, with 80% data going to Model Training and the remaining 20% going to Model Testing.

An 80:20 train-test split was utilized based on a benchmarking or base paper [2] that employed the same ratio to achieve optimal accuracy. This was to be made with the specific aim of coming directly head-to-head with these results and topping them. In addition to that, to increase the level of accuracy even more, the SMOTE algorithm for over-sampling the minority synthetic class was introduced. It was possible to prevent the problem of class imbalance and enhance the performance of the classifier, which proved to be more effective compared to the outcomes presented in the paper under analysis. Cross-validation was not used in order to maintain consistency with the benchmark paper [2].

The 20% testing samples were selected randomly to ensure that the test set was representative of the overall dataset. No clustering techniques were employed for this selection process. The random sampling approach is consistent with the methodology used in the benchmarking paper, allowing for a direct and fair comparison of results and not using clustering techniques.

SMOTE algorithm: The class imbalance problem is addressed by SMOTE this method has been used in ML. This provides better models' performance as it synthesizes through the artificial creation of the minority instances in the data set leading to near-equal class distribution, mostly for the case of the classification tasks.

Figure 5 represents the methodology followed starting from importing the Python packages, loading the dataset into a data frame, performing data preprocessing with and without the SMOTE algorithm, analyzing and validating the dataset, data visualization, scaling data, training and testing data split, model building, model building, model testing, accuracy calculation, accuracy comparison.

The flow of the proposed work is given below.

Step 1 (Import modules):

Importing essential libraries that are needed for the suggested job is the initial step. Graphs are created using the Seaborn and Matplotlib libraries. Training and testing splitter is implemented using the Scikit_learn library. Python's NumPy is used for mathematical and numerical operations, whereas pandas are used for data analysis and manipulation, usually with tabular data.

Step 2 (Loading dataset):

Start Loading the UCI dataset in the Training folder into the RAM during the runtime.

Step 3 (Analyse and validate the dataset):

Using Pandas perform basic operations to know the associated attributes and their data type, to categorize between numerical, and categorical attributes, and to know the categories involved in the categorical attributes and also mainly to know the NULL values if present and if yes then how many and what are those attributes.

Step 4 (Data visualization):

This step is used to know the hidden relationship or pattern between the attributes. Bar plots display the distribution of categorical data; pair plots reveal relationships between pairs of variables; heatmaps visualize correlations or patterns in a dataset, and pie charts represent the composition of a whole as parts.

Step 5 (Scaling data):

Scaling in machine learning adjusts the range or distribution of data, with normalization rescaling to [0, 1] and standardization to mean=0, std=1. It's crucial for ensuring fair feature contribution and better model convergence.

Step 6 (Train–Test split):

Using the Scikit-learn library, we can take advantage of the 'train-test split' feature to partition the dataset cleanly into 2 subsets: the first subset for training and the second subset for testing. This paper presents the findings of both the 80:20 and 70:30 ration splits.

Step 7 (Model building):

Bring in the required model libraries, including DT, KNN, GB, AB, RF, SVM, SGD, and XGB. Assign each of these models to its variable for easy access.

Step 8 (Model Training):

In this stage, the data is passed into the model to train our model with our dataset, where it will find hidden patterns that help it in its prediction of classes for new/unfamiliar inputs.

Step 9 (Testing the model):

Provide input with the independent testing dataset and allow the model to predict target values. In this case, the model utilizes its learning from the training process to predict likely outcomes using the new and previously unseen data.

Step 10 (Accuracy calculation):

Matching the actual world values of the dependent characteristic with those our model anticipated. A tool such as a confusion matrix, can be useful to judge how good or not each model's predictions perform. This enables us to determine the level of success of our model within its practical relevance setting.

Step 11 (Accuracy comparison and best model declaration):

After obtaining the accuracy scores of all the models, then move to the final practical decision. Choose the best one or two exemplary models showing the highest accuracy, and take these as the high-rated models by this dataset. The best-performing models are the ones that will be used in predicting and also drawing meaningful conclusions in the future.

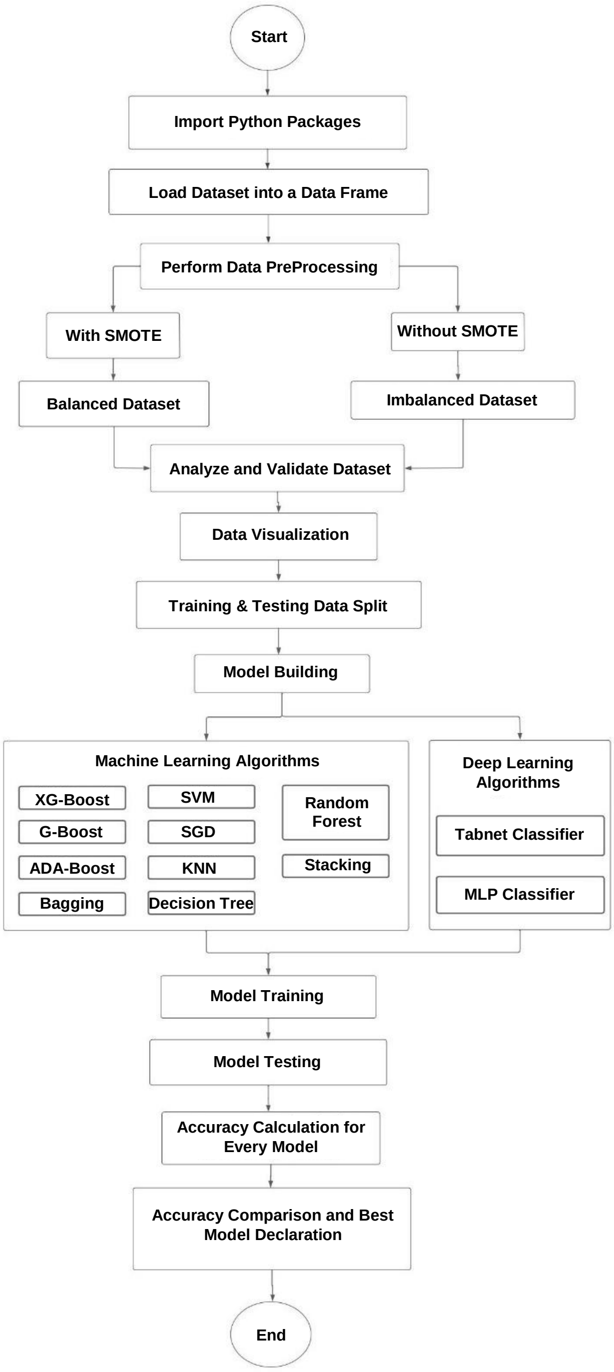

Figure 5 The flowchart for the proposed work

The proposed work uses the accuracy, precision, recall, and F1-Score as performance metrics and TP, TN, FP, and FN are the standard statistical measures often used in binary classification problems and especially in the frame of diagnostic medical examinations or numerical previsions.

True positive (TP) gives a true positive result when there is the presence of a condition or outcome in the individual being tested.

True negative (TN) means that the test results correctly show that the person does not have the condition or outcome being tested for. In this case, all the symptoms implied in the test result are absent, thus the test exists in a negative position, and the condition being tested is truly void.

A false positive (FP) is formally defined as a condition in which someone tests negative for a certain condition but, despite this negative result, is told they have that condition. The daily examination reveals a positive result while the condition being diagnosed is not present.

A false negative (FN) result is characterized by a situation where a test returns negative where in actual sense the condition or outcome is present. As it turns out, there are no problems with the outcome of the test, even though the illness being diagnosed has affected the patient.

The accuracy of a model is measured by dividing the number of correctly predicted occurrences by the total number of instances. The formula for accuracy is shown in the equation.

$$Accuracy = \frac{no.of\ correct\ predictions}{total\ no.of\ instances} \tag{2}$$

The precision of a model relates to its ability to precisely identify important events among all instances that are expected to be positive. It is calculated as the ratio of actual positive forecasts to all expected positives. The formula for precision is shown in the equation 2.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{3}$$

Recall measures a model's ability to detect all relevant occurrences; it is also known as sensitivity or true positive rate. It is calculated as the ratio of true positive forecasts to all actual positives. The formula for the recall is shown in the equation 3.

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{4}$$

The F1-Score is the harmonic mean of recall and accuracy, and it strikes a balance between the two measurements. It is especially useful when the distribution of classes is not uniform. The formula for F1-score is shown in the equation 4.

$$F1 - Score = 2 * \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

4. Results and discussion

Various ML algorithms and DL algorithms are used for obesity prediction studies, including XGB, GB, RF, DT, SVM, KNN, SGD, Stacking, Bagging, AB and TabNet classifier, and MLP classifier, also the dataset is used in two ways: one unbalanced version and one balanced version using SMOTE algorithm. Graphs were also displayed, and it showed that the models with a balanced data set had higher accuracies.

Table 2 shows the parameters followed by each model and the parameters include random state, epochs, patience, batch size, virtual batch size, number of workers, activation function, solver, max iteration, and random seed that are tuned for best accuracy results.

Table 2 Parameters applied to various models

Name of the Algorithm	Parameters used
XGB, GB, RF, DT, SVM, SGD, Stacking, Bagging, AB	Random_state =38
TabNet classifier	epochs =150, patience =150, batch size =64, virtual batch size =32, num of workers=0, random seed = 38
MLP classifier	activation = relu, solver = Adam, max iteration =150, Random_state =38
KNN	NA

According to the study in the publication, the usage of ML algorithms such as XGB & GB and various others has been understood and referred to in works [2-4, 6, 15, 26].

In Table 3, when handling an unbalanced dataset without applying the SMOTE algorithm for obesity, there is a fluctuation in accuracy. Various machine learning techniques and deep learning techniques, like XGB, GB, TabNet classifier, MLP classifier, RF, DT, and others were employed. It is worth noting that the TabNet classifier (96.6%), MLP classifier (94.5%), XGB (96.2%), GB (95.5%), and RF (94.8%) achieved the levels of accuracy among these methods.

Table 3 Accuracy results comparison for imbalanced dataset

Models	Accuracy	Precision	Recall	F1-score	AUC-RUC
XGB	96.2 %	96%	96%	96%	99.9%
GB	95.5 %	96%	96%	96%	99.6%
TabNet	96.6 %	97%	97%	97%	100%
MLP	94.5 %	95%	95%	95%	98%
RF	94.8 %	95%	95%	95%	99.8%
DT	93.6 %	94%	94%	94%	94.9%
SVM	82.5 %	83%	83%	82%	98.0%
KNN	75.9 %	79%	80%	79%	94.8%
SGD	68.6 %	79%	80%	79%	94.3%
Stacking	16.8 %	19%	18%	17%	52%
Bagging	94.07 %	95%	95%	95%	99.1%
AB	28.4 %	79%	80%	79%	78.3%

Figure 6 shows the accuracies of the various models without using the SMOTE algorithm and the knowledge to be inferred from this is that TabNet classifier outperformed all the models followed by XGB, GB, RF, MLP, Bagging, DT, SVM, KNN, SGD, AB, and Stacking.

Figure 7 displays ROC (Receiver Operating Characteristic) curves for various machine learning models: Models such as XGBoost, Gradient Boosting, TabNet, MLP, Random Forest, and Decision Tree without SMOTE Algorithm. This plot represents the accuracy of the model and as can be seen from the AUC figures indicated most classes are very correct.

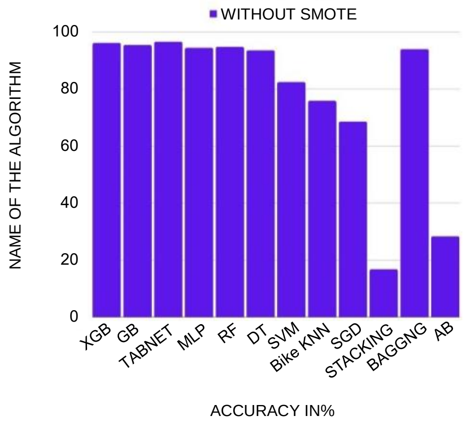

Figure 6 Accuracies of the various models without using the SMOTE algorithm.

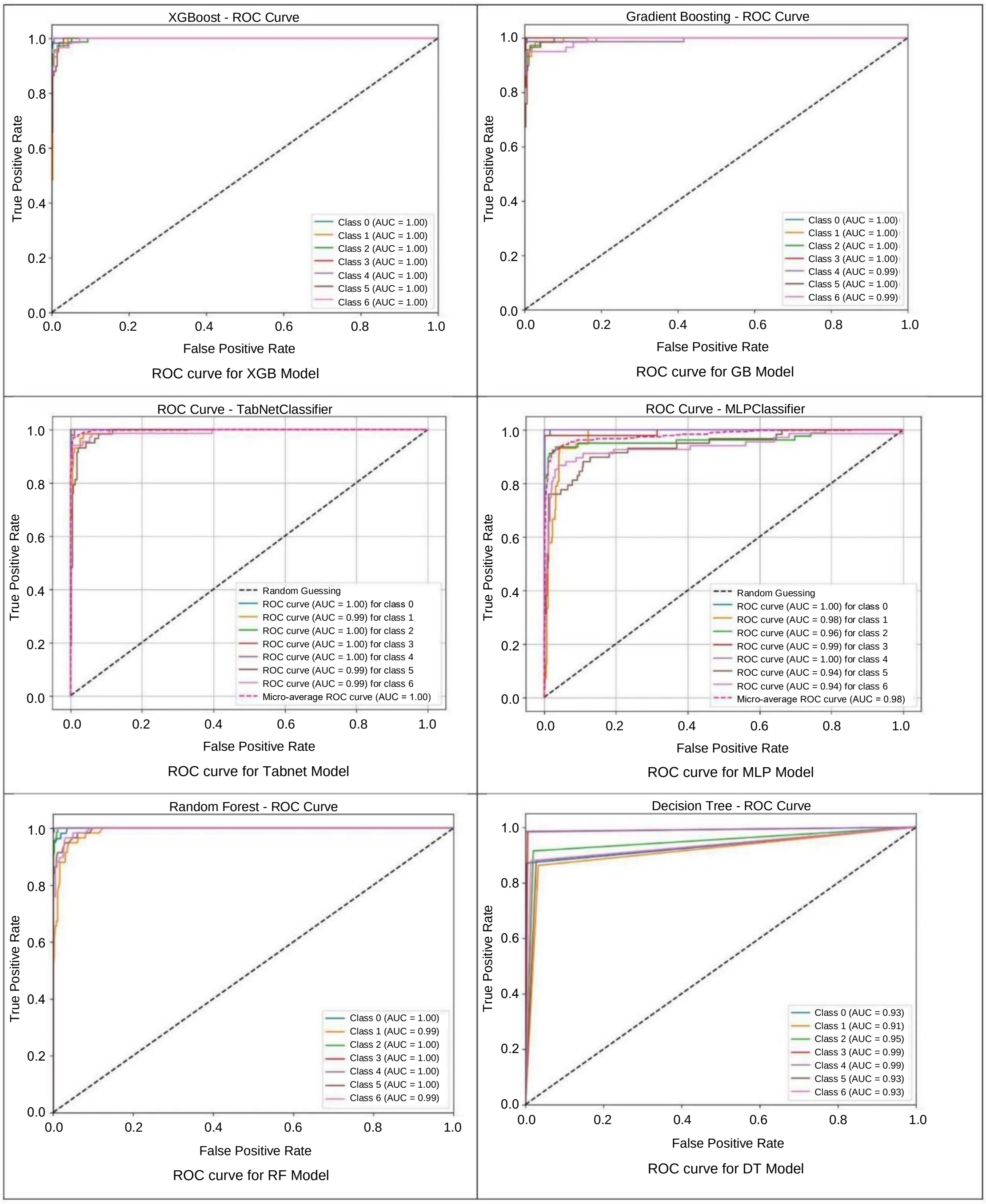

Figure 7 ROC curves for XGB, GB, TabNet, MLP, RF, and DT models without SMOTE Algorithm.

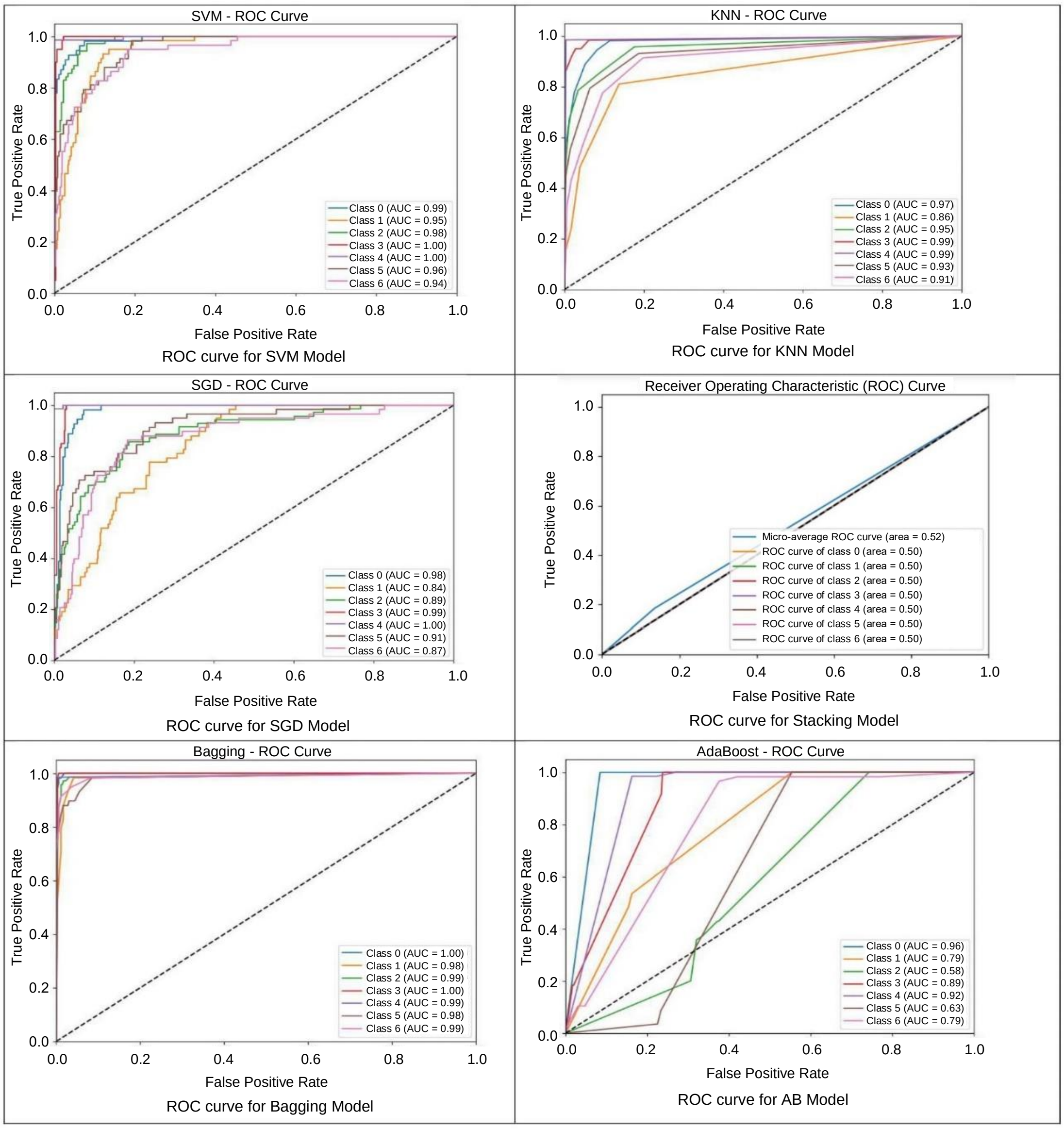

Figure 8 ROC curves for SVM, KNN, SGD, Stacking, Bagging, AB models without SMOTE Algorithm

Figure 8 shows ROC curves for several machine learning models: SVM, KNN, SGD, Stacking, Bagging, and AdaBoost applied without the SMOTE Algorithm. Each plot illustrates the model's performance across multiple classes, with AUC values indicating varying degrees of accuracy, highlighting the strengths and weaknesses of each model in handling different classes.

In Table 4, the performance measures like Precision, Recall, F1-score, and Accuracy of the various classifiers are compared when dealing with a balanced dataset using the SMOTE algorithm for obesity. Different ML techniques and DL techniques, like XGB, GB, TabNet classifier, MLP classifier, RF, DT, and others were employed. It is worth noting that GB (99.3%), XGB (99%), TabNet classifier (98.4%), MLP classifier (97.7%), and RF (97.4%) achieved the levels of accuracy among these methods.

Figure 9 represents the bar plot for all the accuracies of models after applying the SMOTE algorithm and the knowledge to be inferred is that the GB algorithm outperformed all the models followed by XGB, TabNet classifier, MLP classifier, RF, Bagging, DT, SVM, KNN, SGD, AB, Stacking.

Table 4 SMOTE algorithm comparison of accuracy results for a balanced dataset

Models	Accuracy	Precision	Recall	F1-score	AUC-ROC
XGB	99 %	99%	99%	99%	100%
GB	99.3 %	99%	99%	99%	100%
TabNet	98.4 %	98%	98%	98%	100%
MLP	97. 7 %	98%	98%	98%	99%
RF	97.4 %	98%	97%	97%	99.9%
DT	96. 9 %	97%	97%	97%	98%
SVM	90.3 %	91%	90%	90%	99.1%
KNN	85.7 %	90%	90%	90%	96.2%
SGD	67.6 %	90%	90%	90%	93.7%
Stacking	12.30 %	12.3%	12.3%	12.3%	49%
Bagging	96.28 %	96.28%	96.28%	96.28%	99.9%
AB	34.9 %	90%	90%	90%	75.9%

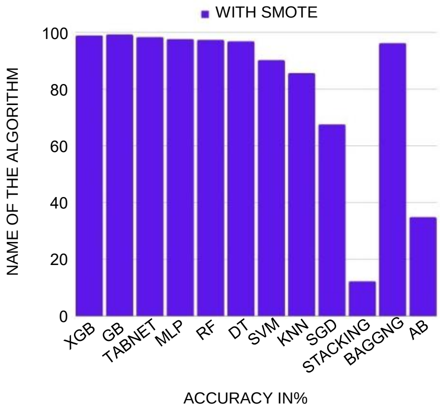

Figure 9 Accuracies of the various models using the SMOTE algorithm.

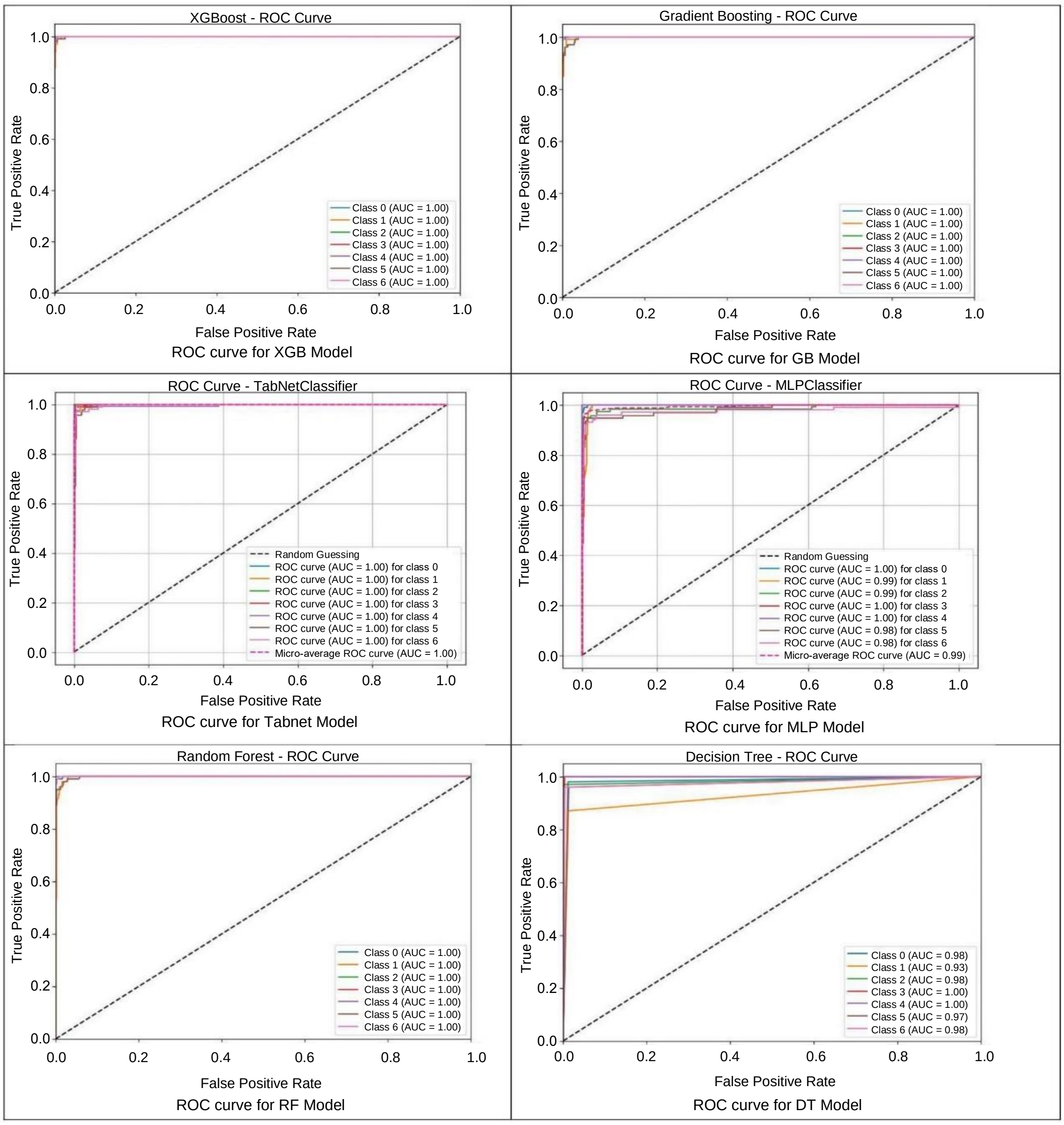

Figure 10 ROC curves for XGB, GB, TabNet, MLP, RF, and DT models with SMOTE Algorithm.

Figure 10 presents ROC curves for various machine learning models: XGBoost, Gradient Boosting, TabNet, MLP, Random Forest, and Decision Tree algorithms and uses a data set applying the SMOTE Algorithm. All the plots highlight the AUC score for the models in the form of graphs and the values recorded are high proving that the proposed models yield higher accuracy and better class distinction for every class.

Figure 11 shows ROC curves for several machine learning models: SVM, KNN, SGD, Stacking, Bagging, and AdaBoost applied with the SMOTE Algorithm. Each plot illustrates the model's performance across multiple classes, with AUC values indicating varying degrees of accuracy, highlighting the strengths and weaknesses of each model in handling different classes.

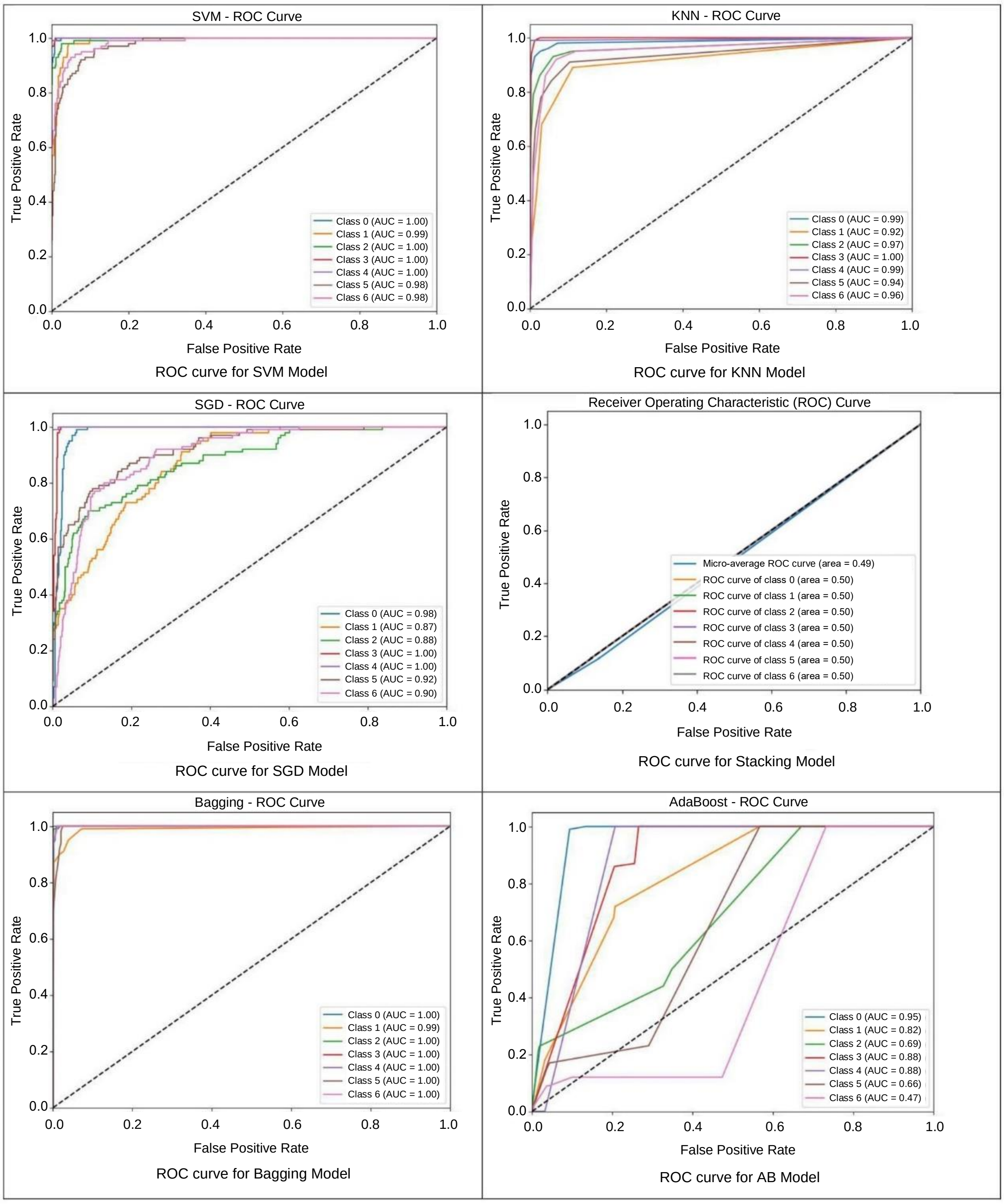

Figure 11 ROC curves for SVM, KNN, SGD, Stacking, Bagging, and AB models With SMOTE Algorithm.

Table 5 compares the accuracy levels of the models before and after applying the SMOTE algorithm, and out of 12 models used, 10 models' accuracies are increased after applying the SMOTE algorithm to the workflow. So, the purpose of the SMOTE algorithm, which is used to balance the dataset, significantly enhanced the proposed model's quality, and the accuracies after the SMOTE algorithm are considered to be final and accepted for this system.

Table 5 Evaluating the dataset's accuracy using several classifiers and both the presence and absence of the SMOTE algorithm used

Models	Model Accuracy without SMOTE algorithm	Model Accuracy with SMOTE algorithm
XGB	96.2 %	99 %
GB	95.5 %	99.3 %
TabNet	96.6 %	98.4 %
MLP	94.5 %	97.7 %
RF	94.8 %	97.4 %
DT	93.6 %	96.9 %
SVM	82.5 %	90.3 %
KNN	75.9 %	85.7 %
SGD	68.6 %	67.6 %
Stacking	16.8 %	12.30 %
Bagging	94.07 %	96.28 %
AB	28.4 %	34.9 %

Figure 12 represents the bar plot for all the model accuracies and gives you a comparative result for with and without SMOTE algorithm data and the knowledge to be inferred is that 10 out of 12 models' accuracies have been increased after applying SMOTE algorithm.

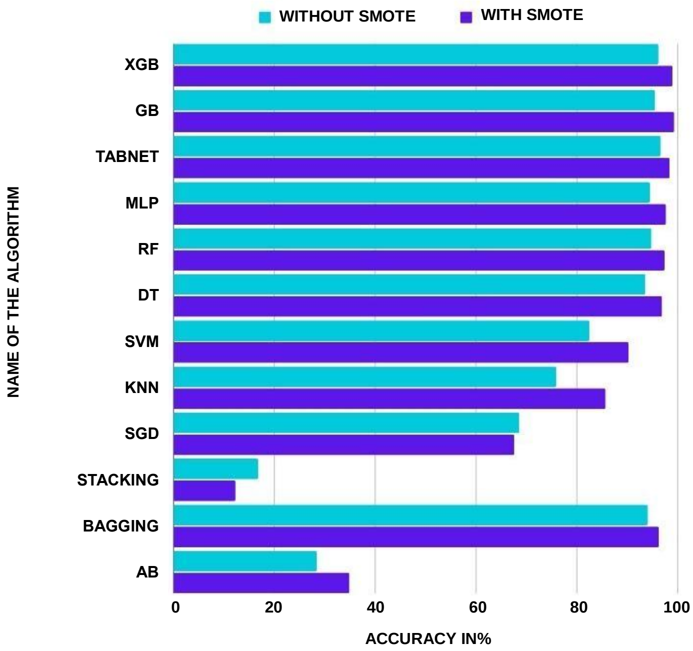

Figure 12 Accuracy comparison of various models with and without using SMOTE algorithm

Table 6 compares the accuracies of existing and proposed models and out of 6 existing models, 4 proposed models have outperformed their accuracy results. They are, GB with 99.3%, RF with 97.4%, DT with 96.9%, AB with 34.9%.

Table 6 Accuracy comparison between existing and proposed model accuracies

Models	Existing model accuracies	Proposed model accuracies
XGB	-	99 %
GB	97.2 %	99.3 %
TabNet classifier	-	98.4 %
MLP classifier	-	97.7 %
RF	94.8 %	97.4 %
DT	96.7 %	96.9 %
KNN	86.3 %	85.7 %
SGD	-	67.6 %
Stacking	-	12.30 %
Bagging	-	96.28 %
AB	33.1%	34.9 %
SVM	43.4%	90.3%

In the existing model, the highest accuracy is achieved by GB with 97.2%, but in the proposed model, 5 out of 12 models have outperformed that benchmarked accuracy, they are GB with 99.3%, XGB with 99%, TabNet classifier with 98.4%, MLP classifier with 97.7% and RF with 97.4% and Stacking has a low accuracy of 12% which implies that the current model configuration is not effective for the used dataset. This conclusion is made while able to affirm that there are no problems concerning featurization, data integrity, or the training step. The cleaning of the dataset is good, features have been properly selected or extruded, and the training phase has been done in the right manner. Hence, there are main factors behind the poor performance and that is related to the base models and meta-model used in the stacking ensemble. This indicates that the selected algorithms or at least one of them, or their mixture, do not learn the necessary dependencies in the data adequately.

Working with a small dataset also has several disadvantages such as overfit, underfit, limited number of data samples, high variance, problems with the assessment of a model, and low statistical significance. Such limitations result in poor models and increased randomness in the evaluation metrics, which was demonstrated by the low accuracy of the model. These problems and to enhance the model's generalization, the SMOTE, which stands for Synthetic Minority Over-sampling Technique, has been selected to augment the records in a given dataset. Even though it does not introduce completely new objects to the model, SMOTE improves the model by supplementing it with more balanced classes and at the same time prepares the data for getting better results with unseen data.

A ''random state'' is an argument that is commonly used in numerous machine learning models to determine the level of randomness that is allowed when splitting the data. It helps in making the outcomes of the experiment reproducible by setting the seed of the random number generator. In setting the random state as a certain number, the same splitting of data into training and testing sets will be produced, making the results reproducible and comparable.

The testing sample of 20% was not chosen randomly. However, there is a random state parameter present, normally they set it to 0 or 42. In this paper, random state parameter testing was performed at every value from 0 to 42 for every model. By doing so, it was identified that the state value of 38 provided the highest accuracy for nearly all the proposed models. Thus, the random state parameter is set at 38 and used for every model that accepts the parameter and is indicated accordingly. There were no trials of clustering techniques for the splitting process, as the benchmarking paper [2] used mentioned the train-test split with the testing sample comprising 20%.

5. Conclusion and future work

This would imply that GB had an edge over the other considered algorithms in accurately predicting obesity. SMOTE algorithm also helped to balance out class ratios and improve overall model performance. This implies that machine learning models can provide accurate projections. GB recorded higher scores than every other algorithm at a whopping 99.3%.

Additionally, the SMOTE algorithm was vital in balancing the data's class imbalance. When the numbers of instances in one instance exceed significantly the number in the other instances, a situation referred to as class imbalance occurs. The dataset in the obesity prediction might have included fewer obese individuals and more non-obese people. In this study, the SMOTE algorithm is employed for oversampling the minority class and generating artificial data points as well which helped enhance our machine learning model performance. The use of this method helped in curbing bias and minimizing chances whereby the models became biased towards the majority of the class.

Nevertheless, it should be emphasized that these results are preliminary, and more research and verification are required to prove that these algorithms predict obesity. This will provide stronger evidence of how they perform with larger, more varied datasets if experimented with. Moreover, one may employ other approaches that can be used in class imbalance like under-sampling and mixed methods for improvement of enhanced models' accuracy.

Lastly, this paper suggests applying ML models and DL models such as GB and Tabnet classifiers respectively in estimating obesity. Combining both models individually with an algorithm called SMOTE helped to address the imbalance in classes and improved the accuracy of models. This will build upon the growing body of knowledge about predicting obesity, as well as lay the groundwork for further investigation.

6. References

[1] Kivrak M. Deep learning-based prediction of obesity levels according to eating habits and physical condition. J Cogn Syst. 2021; 6(1):24-7.
[2] Maria AS, Sunder R, Kumar RS. Obesity risk prediction using machine learning approach. 2023 International Conference on Networking and Communications (ICNWC); 2023 Apr 5-6; Chennai, India. USA: IEEE; 2023. p. 1-7.
[3] Anisat MF, Basaky FD, Osaghae EO. Obesity prediction model using machine learning techniques. J Appl Artif Intell. 2022;3(1):24-33.
[4] Alsareii SA, Awais M, Alamri AM, AlAsmari MY, Irfan M, Raza M, et al. Machine-Learning-Enabled obesity level prediction through electronic health records. Comput Syst Sci Eng. 2023;46(3):3715-28.
[5] Yagin FH, Gülü M, Gormez Y, Castañeda-Babarro A, Colak C, Greco G, et al. Estimation of obesity levels with a trained neural network approach optimized by the Bayesian technique. Appl Sci. 2023;13(6):3875.
[6] Kaur R, Kumar R, Gupta M. Predicting risk of obesity and meal planning to reduce the obese in adulthood using artificial intelligence. Endocrine. 2022;78(3):458-69.
[7] Quiroz JPS. Estimation of obesity levels based on dietary habits and condition physical using computational intelligence. Inform Med Unlocked. 2022;29:100901.
[8] Cui T, Chen Y, Wang J, Deng H, Huang Y. Estimation of obesity levels based on Decision trees. 2021 International Symposium on Artificial Intelligence and its Application on Media (ISAIAM); 2021 May 21-23; Xi'an, China. USA: IEEE; 2021. p. 160-5.
[9] Celik Y, Guney S, Dengiz B. Obesity level estimation based on machine learning methods and artificial neural networks. 2021 44th International Conference on Telecommunications and Signal Processing (TSP); 2021 Jul 26-28; Brno, Czech Republic. USA: IEEE; 2021. p. 329-32.
[10] Ferdowsy F, Rahi KSA, Jabiullah MI, Habib MT. A machine learning approach for obesity risk prediction. Curr Res Behav Sci. 2021;2:100053.
[11] Jeon J, Lee S, Oh C. Age-Specific risk factors for the prediction of obesity using a machine learning approach. Front Public Health. 2023;10:1-12.
[12] Molina D, De-La-Hoz A, Mendoza F. Classification and features selection method for obesity level prediction. J Theor Appl Inf Technol. 2021;99(11):2525-36.
[13] Rodríguez E, Rodríguez E, Nascimento L, da Silva A, Marins F. Machine learning techniques to predict overweight or obesity. IDDM-2021: 4th International Conference on Informatics & Data-Driven Medicine; 2021 Nov 19-21; Valencia, Spain. p. 190-204.
[14] Pinto KA, Abdullah NL, Keikhosrokiani P. Diet & exercise classification using machine learning to predict obese patient's weight loss. 2021 International Congress of Advanced Technology and Engineering (ICOTEN); 2021 Jul 4-5; Taiz, Yemen. USA: IEEE; 2021. p. 1-5.
[15] Cervantes RC, Palacio UM. Estimation of obesity levels based on computational intelligence. Informa Med Unlocked. 2020;21: 100472.
[16] Deniz S, Şirin H, Kıvrak M, Kaplan Z, Ketrez G, Üner S. Factors associated with overweight and obesity in students of 5-14 age group in Mersin. Gulhane Med J. 2020;62(4):245-53.
[17] De-La-Hoz-Correa E, Mendoza-Palechor FE, De-La-Hoz-Manotas A, Morales Ortega RC, Beatriz Adriana SH. Obesity level estimation software based on Decision trees. J Comput Sci. 2019;15(1):66-77.
[18] 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. 2019;25:104344.
[19] DeGregory KW, Kuiper P, DeSilvio T, Pleuss JD, Miller R, Roginski JW, et al. A review of machine learning in obesity. Obes Rev. 2018;19(5):668-85.

[20] Montañez CAC, Fergus P, Hussain A, Al-Jumeily D, Abdulaimma B, Hind J, et al. Machine learning approaches for the prediction of obesity using publicly available genetic profiles. 2017 International Joint Conference on Neural Networks (IJCNN); 2017 May 14-19; Anchorage, USA. USA: IEEE; 2017. p. 2743-50.
[21] Yahia N, Wang D, Rapley M, Dey R. Assessment of weight status, dietary habits and beliefs, physical activity, and nutritional knowledge among university students. Perspect Public health. 2016;136(4):231-44.
[22] Dugan TM, Mukhopadhyay S, Carroll A, Downs S. Machine learning techniques for prediction of early childhood obesity. Appl Clin Inform. 2015;6(3):506-20.
[23] Davila-Payan C, DeGuzman M, Johnson K, Serban N, Swann J. Estimating prevalence of overweight or obese children and adolescents in small geographic areas using publicly available data. Prev Chronic Dis. 2015;12:1-13.
[24] Muhamad Adnan MHB, Husain W, Abdul Rashid N. A hybrid approach using Naïve Bayes and Genetic Algorithm for childhood obesity prediction. 2012 International Conference on Computer & Information Science (ICCIS); 2012 Jun 12-14; Kuala Lumpur, Malaysia. USA: IEEE; 2012. p. 281-5.
[25] Muhamad Adnan MHB, Husain W, Damanhoori F. A survey on utilization of data mining for childhood obesity prediction. 8th Asia-Pacific Symposium on Information and Telecommunication Technologies; 2010 Jun 15-18; Kuching, Malaysia. USA: IEEE; 2010. p. 1-6.
[26] Calderón-Díaz M, Serey-Castillo LJ, Vallejos-Cuevas EA, Espinoza A, Salas R, Macías-Jiménez MA. Detection of variables for the diagnosis of overweight and obesity in young Chileans using machine learning techniques. Procedia Comput Sci. 2023;220:978-83.
[27] Pereira NC, D'souza J, Rana P, Solaskar S. Obesity-related disease prediction from healthcare communities using machine learning. 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT); 2019 Jul 6-8; Kanpur, India. USA: IEEE; 2019. p. 1-7.
[28] Nicklas TA, Yang SJ, Baranowski T, Zakeri I, Berenson G. Eating patterns and obesity in children. The Bogalusa Heart Study. Am J Prev Med. 2003;25(1):9-16.
[29] Zhao R, Zhao L, Gao X, Yang F, Yang Y, Fang H, Ju L, et al. Geographic variations in dietary patterns and their associations with overweight/obesity and hypertension in China: findings from China nutrition and health surveillance (2015–2017). Nutrients. 2022;14(19):3949.
[30] Alkhalaf M, Yu P, Shen J, Deng C. A review of the application of machine learning in adult obesity studies. Appl Comput Intell. 2022;2(1):32-48.
[31] Devi KN, Krishnamoorthy N, Jayanthi P, Karthi S, Karthik T, Kiranbharath K. Machine learning based adult obesity prediction. 2022 International Conference on Computer Communication and Informatics (ICCCI); 2022 Jan 25-27; Coimbatore, India. USA: IEEE; 2002. p. 1-5.