Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Machine Learning Modelling for Imbalanced Dataset: Case Study of Adolescent Obesity in Malaysia

Nur Liana Ab Majid[1,*], Syahid Anuar[2]

[1] Institute for Public Health, National Institutes of Health, Ministry of Health Malaysia, Seksyen U13, Bandar Setia Aam, 40170 Shah Alam, Selangor, Malaysia
[2] Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 23 August 2023
Received in revised form 3 October 2023
Accepted 4 November 2023
Available online 25 December 2023

Keywords:
Prediction; Adolescent obesity; Imbalanced dataset

ABSTRACT

Obesity among adolescent is a public health issue with increasing burden of disease. Predicting imbalanced health data with Machine Learning may introduce bias and lead to diminished model performance. Misclassification in healthcare data could lead to misdiagnosing a patient or failing to detect a health issue when it is present. The purpose of this study is to predict adolescent obesity using machine learning along with implementation of multiple approaches on the imbalanced dataset. This study used secondary dataset from National Health and Morbidity Survey 2017. Samples 13 – 17 years were selected for the classification. SPSS V26 was used for data pre-processing, data cleaning, and data analysis. Meanwhile, Python language used for prediction and evaluation of the models. Approaches on the imbalanced dataset including resampling method (Random Oversampling, Random Under-sampling) and hybrid method (SMOTE and ADASYN) were implemented. This dataset was used for the formation of predictive models on ML algorithm including Artificial Neural Network, Decision Tree, K-Nearest Neighbour, Logistic Regression, Naïve Bayes, Random Forest and Support Vector Machine. The performance of each model was evaluated and compared using accuracy, precision, recall, F- score and Area under the Curve (AUC). Random Oversampling approached with Decision Tree Algorithm performs the best with accuracy (91.35%), precision (0.93), recall (0.91), F- score (0.91) and AUC (0.91) for the prediction of obesity among adolescent in Malaysia. The presented ML model development workflow along with the imbalanced techniques can be adapted to other health survey-based studies and may be valuable for developing other clinical prediction models.

1. Introduction

The advantages of Machine Learning (ML) over statistical methods have attracted many interests for it to be used in the area of medical diagnoses [1,2]. The capacity of ML to deal with high-dimensional data and its ability to find complex, nonlinear relationships in an automated way far outclasses those of the traditional statistical models. As obesity researchers and healthcare professionals have access to a wealth of data, the trends of machine learning applications in the

* *Corresponding author.*
E-mail address: nurlianaabmajid@gmail.com

https://doi.org/10.37934/araset.36.1.189202

obesity research field have been increasing simultaneously. With the availability of large datasets, machine learning provides sophisticated and elegant tools to describe, classify and predict obesity related risks and outcomes. To prevent childhood and adolescent obesity, it is of considerable value to build predictive models that can identify those at high risk sooner. This permits the concentration of preventative efforts on the high-risk subgroup, enabling a more cost-effective and individualized approach to weight loss programmes.

Childhood overweight and obesity often persist into adulthood and thus increase the risks of developing non-communicable diseases such as diabetes and cardiovascular risk at an earlier age [3], as well as to a lower quality of life [4]. Over the past 40 years, there has been a significant increase in the prevalence of obesity and overweight among children and adolescents [5]. Between 1975 and 2016, there were 18% more overweight and obese children and adolescents worldwide between the ages of 5 and 19. The difficulties in eradicating the disease once it has become entrenched justifies the adoption of preventative rather than curative interventions for children [6]. The global pooled prevalence for obesity in children aged 5–11 years and in adolescents aged 12–19 years were 5.8% and 8.6% respectively [7]. Data from the National Health and Nutrition Examination Survey (NHANES) 2009 – 2010 indicated a prevalence of overweight and obesity among adolescents aged 12 – 19 years in the United States were 15.2% and 18.4% respectively. Meanwhile, in Africa the obesity prevalence among adolescents 13 – 19 years were 5.3%, China (13 – 17 years) at 0.6% and Saudi Arabia (13 – 18 years) at 7.0% [8]. In Malaysia, the prevalence of obesity in children and adolescents aged 5 to 17 years were 14.8% [9].

The occurrence of obesity among the adolescents are low compared to the overall population, resulting in an imbalanced distribution of classes. This unequal distribution leads to data balancing issues which may affect the outcome of the predictive mode, producing high false negative results. The term "imbalance problem" refers to the occurrences of one of the classes, the majority class, are much larger than those of the other class, the minority class, or that the number of instances of the majority class exceed the number of instances of the minority class [10]. Adolescent obesity is also called the minority class due to the smaller prevalence compared to the other majority class or normal class. Study by Singh and Tawfik [11], conducted machine learning for adolescent obesity classification and found out that the accuracy for the minority class to be very poor. Despite having vast dataset on obesity, the most common issue related to medical diagnoses using machine learning is its performance when utilizes imbalanced dataset.

The imbalanced data sets problem plays a key role in machine learning. Predictions related to the disease displayed a bias favouring the majority class, which consequently eroded the trustworthiness of machine learning models. This bias could have led to misdiagnoses and mismanagement of patients, and in the worst-case scenario, even fatal outcomes. Hence, it is crucial to prioritize and enhance the reliability of the existing ML models for disease occurrence prediction. This issues was highlighted by Singh and Tawfik [11] and Pranto *et al.,* [12] in their studies along with the significant improvements on the evaluation of the classification models using balance dataset.

This study aims to implement machine learning for classification of adolescent obesity with approaches on the imbalanced dataset. The performance of imbalanced classification will be evaluated and measured using the selected evaluation matrix. This article is organized into four sections. The next section elaborates on the related works on imbalanced methods used in medical diagnoses and for the prediction of obesity. Section 3 deals with the methodology, the tools and techniques that were employed in this study. Section 4 will present the findings on the implementation of multiple approaches for handling imbalanced dataset, evaluation and comparison of the model performance. The research is summarized and concluded in Section 5 with some recommendations and future work proposals.

2. Related Works

There have been a growing number of machine learning applications in the field of medicine as well as obesity research [10,11,13]. Additionally, the imbalanced data issue has attracted a great deal of attention, and numerous scholars have made contributions to this topic. In this part, the associated works on the unbalanced dataset will emphasize the medical diagnosis made by other researchers. In imbalanced data classification, the proportional class sizes of a dataset diverge by a large margin. The minority class is represented by a small number of samples and the majority class comprises the remainder. As a result, the performance of a classifier in an imbalanced data set tends to favour the majority class. Performance bias indicates that solutions behave differently for majorities and minorities. Solutions in the majority class tend towards greater precision. On the minority class side, the solutions are executed with inadequate precision [10].

Three primary strategies for addressing imbalance issues in data classification were identified including the pre-processing methods, algorithmic oriented approaches, and hybrid methods [10]. Among the literatures, it has been demonstrated that data resampling family procedures may be beneficial and straightforward in the healthcare setting [16,17]. Data resampling can be used to create more even class distributions before training a classifier, by either oversampling the underrepresented class or Under-sampling the overrepresented class. Under-sampling may lose potentially helpful information for accurate classification, while oversampling may lead to overfitting the minority class [10].

There were many studies that proposed the implementation of oversampling methods for the prediction of heart disease, diabetes and obesity using the most common approach – the Synthetic Minority Oversampling Technique (SMOTE) [10,12,18,19]. SMOTE is a famous approach used for the construction of a classifier for the imbalance dataset. When working with an unbalanced dataset, the SMOTE preprocessing approach is among the most trusted available. Since its inception, many iterations of SMOTE have been developed and implemented to make the present SMOTE method more robust and flexible.

There were also study by Kokkotis *et al*., [20] suggested the implementation of random under sampling (RUS) for prediction of stroke. The findings demonstrated that RUS produced a low false-negative rate and a satisfactory Geometric Mean (G-Mean) score. Other study by Varotto *et al*., [21] employs oversampling, under-sampling, and ensemble classification to evaluate a variety of resampling strategies for the unbalanced domain in brain neural recordings from surgically treated patients with focal epilepsies. In comparison to the initial data set, the resampling methods performed better and were shown to be more sensitive to the classification method used, with the best results coming from Adaptive Synthetic Sampling (ADASYN).

Similar study by Oskouei and Bigham [22] focused on re-sampling methods in unsupervised learning on 13 real datasets. The study did both over and under- sampling of the dataset and used J48 Decision Tree and Naïve Bayes as classifiers. Oversampling includes SMOTE, Borderline SMOTE and under sampling includes One Side Selection (OSS) and Neighborhood Cleaning Rule (NCL). The methods were compared with two accuracy criteria True Positive rate (TPrate) and True Negative rate (TNrate). From the study, the oversampling methods run better on the strongly imbalanced datasets, and there is no meaningful difference between two methods on the normal imbalanced datasets. The classifier showed that J48 Decision Tree had higher accuracy ratio in comparison with Naïve Bayes (NB). The evaluation metrics commonly used for the imbalanced classification include F1 score, accuracy, precision, recall and AUC [12,23].

There were multiple studies found on the obesity classification involving adults and youth. Among the algorithms used were Multi-layer Perceptron Feed Forward Artificial Neural Networks

(MLPFANN) [24], Naïve Bayes, Random Tree [25], XGBoost, Random Forest (RF) and SVM [13]. The imbalanced data issues were also raised in some of the study including SMOTE [11,26], ADASYN [18] and the ensemble method [14]. Literature searches revealed a limited amount of research undertaken to predict childhood obesity. This might be a result of the problem's complexity. Among the algorithms which performs the best were Bayesian Network (BN), Decision Tree (DT J48), Naïve Bayes (NB), Artificial Neural Network (ANN), Support Vector Machine (SVM) [27], ID3 [28] and XGBoost algorithm [29].

In general, multiple ML algorithms were applied for obesity classification. These algorithms include Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forests (RF), Decision Trees (DT), k-Nearest Neighbors (KNN), and Naive Bayes (NB). ANN, inspired by neural networks in the human brain, can tackle various tasks like classification and image recognition. It mimics the structure of the human brain with interconnected nodes and is versatile for various tasks. Support Vector Machines (SVM) are a powerful classification and regression algorithm that finds a hyperplane to separate data points. Decision Trees create a tree-like structure for interpretable models and decision-making, while Naive Bayes employs Bayes' theorem for probabilistic classification. Random Forest is an ensemble of decision trees to reduce overfitting [18]. K-Nearest Neighbors (KNN) makes predictions based on neighbouring data points, and Logistic Regression models binary outcomes' probabilities. A detailed explanation of the machine learning algorithms was provided in Colmenarejo's study [30]. Subsequently, most studies evaluated the performance of the models using accuracy, sensitivity, specificity and AUC [13,26]. Meanwhile, some other studies measures the performance using the precision, recall metric and F1 score [11,14,18].

3. Methodology

3.1 Operational Framework and Data Preparation for Obesity Classification

The processes involved in this study were explained in the form of the operational framework (Figure 1). This study used the secondary dataset from the National Health and Morbidity Survey: Adolescent Health Survey (AHS) 2017. It was obtained and requested from the National Institutes of Health Malaysia and this study was registered under the National Medical Research Register (NMRR) [Research ID: RSCH ID-22-04523-WAY] and Medical Research & Ethics Committee (MREC), Ministry of Health Malaysia.

The dataset was pre-processed to ensure that the data used for the modelling is accurate, relevant and complete. Data was cleaned to deal with missing data, outliers and removal of null values. The individual datasets underwent a series of cleaning steps, which included renaming headers, restructuring data, relabelling categorical groupings, merging or removing categories in categorical data, changing data types, and eliminating uncommon columns. Subsequently, missing values, unclassified data, and outliers were identified and addressed by either removing or replacing them to ensure the dataset's completeness and relevance for model development.

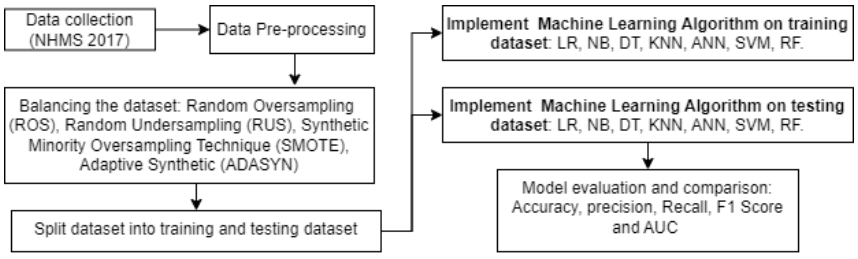

Fig. 1. Flow diagram of research work

The original sample from Adolescent Health Survey (AHS) 2017 consisted of respondents aged between 10 and 18 years old, with a total of 27,497 participants. The dataset contains 261 variables, encompassing a wide range of information collected from the respondents. The samples for prediction of adolescent obesity utilized data from all adolescents aged 13 to 17 years. The variables related to the prediction of adolescent obesity were identified. Among the variables are sociodemographic variables, lifestyle behaviour, dietary intake and substance use. The list of the variables is listed in Table 1. Following the preprocessing steps, a clean dataset was produced, consisting of a total of 26,164 samples and 24 variables including the target outcome.

Table 1
Variables for prediction of adolescent obesity from Adolescent Health Survey 2017 dataset

Variables	Description	Type of variables
Sociodemographic		
Gender	Male, Female	Categorical
Age	13, 14, 15, 16, 17 years	Continuous
Locality	Urban, Rural	Categorical
Ethnicity	Malay, Chinese, Indian, Bumiputra Sabah, Bumiputra Sarawak, Other Ethnicities	Categorical
Parents marital status	Married and living together, married but living apart, Divorced, Widower, Separated, do not know	Categorical
Lifestyle behaviour		
Days physically active	0, 1, 2, 3, 4, 5, 6, 7 days	Continuous
Days walking or cycling to or from school	0, 1, 2, 3, 4, 5, 6, 7 days	Continuous
Hours of sitting	Less than 1 hour per day, 1 to 2 hours per day, 3 to 4 hours per day, 5 to 6 hours per day, 7 to 8 hours per day, more than 8 hours per day	Categorical
Ever had sex	Yes, No	Categorical
Internet Addiction	Yes, No	Categorical
Substance abuse		
Second-hand smoker	Yes, No	Categorical
Current cigarette smoking	Yes, No	Categorical
Current e- cigarette smoking	Yes, No	Categorical
Current alcohol intake	Yes, No	Categorical
Current drug intake	Yes, No	Categorical
Current marijuana intake	Yes, No	Categorical
Dietary practice		
Food insecurity	Never, Rarely, Sometimes, Most of the time, Always	Categorical
Vegetables intake	Did not eat vegetables, Less than 1 time per day, 1 time per day, 2 times per day, 3 times per day, 4 times per day, 5 or more times per day	Categorical
Fruits intake	Did not eat fruit, Less than 1 time per day, 1 time per day, 2 times per day, 3 times per day, 4 times per day, 5 or more times per day	Categorical
Fast-food intake	0, 1, 2, 3, 4, 5, 6, 7 days	Continuous
Milk products intake	Did not drink milk or eat milk products, Less than 1 time per day, 1 time per day, 2 times per day, 3 times per day, 4 times per day, 5 or more times per day	Categorical
Plain water intake	Did not drink plain water, Less than 1 time per day, 1 time per day, 2 times per day, 3 times per day, 4 times per day, 5 or more times per day	Categorical

Carbonated soft drinks intake	Did not drink carbonated drink, Less than 1 time per day, 1 time per day, 2 times per day, 3 times per day, 4 times per day, 5 or more times per day	Categorical
Obesity Status (Target)	0 – non obesity, 1 - obesity,	

3.2 Data Balancing, Modelling and Evaluation

In this study, the obesity prediction was considered into two- class classification, the non- obesity and obesity classes. After preprocessing the dataset, the next step involved addressing class imbalance through various techniques like Random Oversampling (ROS), Random Undersampling (RUS), Synthetic Minority Over-sampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN). Random Oversampling (ROS) randomly replicates minority class instances to achieve a more balanced dataset. Meanwhile, Random under-sampling, randomly selected and reduced a subset of majority class instances to balance the dataset. More advanced methods of oversampling like Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic (ADASYN) were used to generate synthetic minority class samples. *Imblearn* library was used for the implementation of the techniques to deal with the imbalanced dataset so that an equal quantity of non- obesity and obesity classes can be obtained.

Subsequently, the dataset was split into training and testing sets, followed by the application of machine learning algorithms on both datasets. Multiple ML algorithms used for obesity classification namely Artificial Neural Network, Decision Tree, K-Nearest Neighbour, Logistic Regression, Naïve Bayes, Random Forest and Support Vector Machine. Finally, the model's performance was evaluated using metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC).

Accuracy is the proportion of correct predictions made by the model out of the total predictions, which involves distinguishing between obese and non-obese cases. It provides an overall measure of the model's correctness [26]. Precision is the ratio of true positive predictions to the sum of true positive and false positive predictions. In this study, precision is the proportion of true obese samples being classified among all samples classified as obese. Precision measures the model's ability to correctly identify positive instances. Recall, also referred to as sensitivity or true positive rate, quantifies the ratio of true positive predictions to the total of true positive and false negative predictions. In this study, recall is the proportion of capturing obese samples and has a low rate of missing or falsely classifying obese samples as negatives. It assesses the model's effectiveness in accurately capturing positive or obesity instances. The F1 score is a balanced measure of the model's accuracy, as it represents the harmonic mean of precision and recall. The (AUC) represents degree or measure of separability that is how much model can differentiate between the classes. It assesses the model's capacity to distinguish between obese and non-obese instances at varying classification thresholds. The definition of the measurements were also mention in other studies [30-32]. The formula for each measurement as shown in Figure 2.

The entire procedure was conducted using the Python programming language, with the assistance of the scikit-learn (sklearn) library. Selected model after being evaluated will be implemented on separated testing NHMS data to see the result of prediction. *Google Colaboratory* is used as a development environment to aid in python-based development.

$$Accuracy = \frac{true\ positive\ (tp) + true\ negative\ (tn)}{true\ positive\ (tp) + true\ negative\ (tn) + false\ positive\ (fp) + false\ negative\ (fn)}$$

$$Precision = \frac{true\ positive\ (tp)}{true\ positive\ (tp) + false\ positive\ (fp)}$$

$$Recall = \frac{true\ positive\ (tp)}{true\ positive\ (tp) + false\ negative\ (fn)}$$

$$f1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$AUC = \frac{1}{2} \times (Sensitivity + Specificity)$$

Fig. 2. Formula for evaluation measurement

4. Results and Discussion

In the AHS 2017 data, 86.52% (n = 22,638) of adolescents were categorized as obese, while 13.48% (n = 3,526) fell into the non-obese category (Figure 3). The original dataset labelled as Imbalance Dataset consist of 22,638 non-obese samples and 3,526 obese samples. When applying the imbalanced methods, Random Oversampling and SMOTE consist of equally non-obese and obese samples of 22,638. Random under-sampling consists of 3,526 samples for both groups. When using ADASYN, there are minor difference between samples for non-obese (22,638) and obese (22,128).

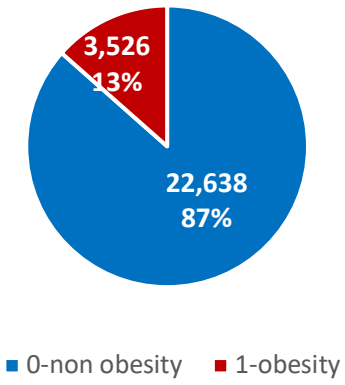

Fig. 3. Percentage of obesity and non-obesity among the adolescent

The number of samples after implementation of the method as listed in Figure 4. The dataset was then divided into 80% training and 20% testing sets.

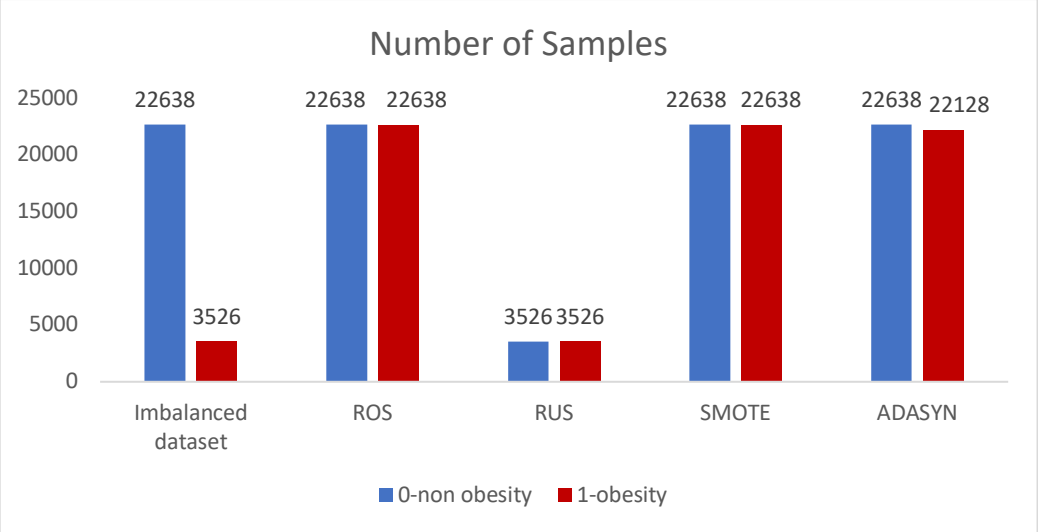

Fig. 4. Number of samples in imbalanced dataset and balanced dataset

The number of samples in testing and training dataset based on imbalanced dataset method as listed in Table 2.

Table 2
Splitting of samples dataset for training and testing

Imbalanced Dataset Methods	Total Sample	Training (80%)	Testing (20%)
Imbalanced Dataset	26,164	20,931	5,233
Random Oversampling	45,276	36,220	9,056
Random Under-sampling	7,052	5,641	1,411
SMOTE	45,276	36,220	9,056
ADASYN	44,766	35,812	8,954

4.1 Model Performance Evaluation

When using imbalanced dataset, the accuracy of imbalanced dataset range between 74 – 86% for LR, KNN, DT, NB and ANN algorithms (Table 3). The accuracy reduced when other method for class imbalanced was introduced. The highest accuracy was achieved using DT modelling along with the ROS technique (91.35%). On the other hand, ANN, RF and SVM consistently exhibit low accuracy across the various approaches used.

Table 3
Accuracy of the obesity classification modelling

Algorithm	Imbalanced	RUS	ROS	SMOTE	ADASYN
Logistic Regression	86.40	56.27	58.34	71.55	70.67
K-Nearest Neighbour	84.80	50.46	78.04	78.37	78.61
Decision Tree	74.22	52.45	91.35	78.26	77.36
Naïve Bayes	86.26	55.92	55.31	61.88	60.05
Artificial Neural Network	86.41	49.18	49.79	49.80	51.08
Random Forest	13.59	49.18	49.79	49.79	48.92
Support Vector Machine	13.59	49.18	49.79	49.79	48.92

In terms of precision, LR, KNN, DT, NB, and ANN algorithms demonstrated moderate performance, with precision values ranging from 0.75 to 0.79 when using imbalanced dataset.

However, the RUS technique yields poor results, while other imbalanced classification methods achieve moderate performance. Notably, KNN performs well when ROS (0.81), SMOTE (0.84), and ADASYN (0.84) were applied. DT achieves the highest precision when using ROS (0.93). On the other hand, ANN, RF and SVM consistently exhibit low precision across all evaluated methods (Table 4).

Table 4
Precision of the obesity classification modelling

Algorithm	Imbalanced	RUS	ROS	SMOTE	ADASYN
Logistic Regression	0.75	0.56	0.58	0.72	0.71
K-Nearest Neighbour	0.76	0.50	0.81	0.84	0.84
Decision Tree	0.77	0.52	0.93	0.79	0.78
Naïve Bayes	0.79	0.56	0.56	0.68	0.66
Artificial Neural Network	0.75	0.24	0.25	0.25	0.26
Random Forest	0.02	0.24	0.25	0.25	0.24
Support Vector Machine	0.02	0.24	0.25	0.25	0.24

In terms of recall, LR, KNN, DT, NB, and ANN algorithms using imbalanced dataset demonstrated good performance, with precision values ranging from 0.74 to 0.86. However, the RUS technique yielded poor results, while other imbalanced classification methods achieved poor to moderate recall value. DT achieved the highest recall when using ROS (0.91). Similar with accuracy and precision, ANN, RF and SVM consistently exhibit low recall across all evaluation methods (Table 5).

Table 5
Recall of the obesity classification modelling

Algorithm	Imbalanced	RUS	ROS	SMOTE	ADASYN
Logistic Regression	0.86	0.56	0.58	0.72	0.71
K-Nearest Neighbour	0.85	0.50	0.78	0.78	0.79
Decision Tree	0.74	0.52	0.91	0.78	0.77
Naïve Bayes	0.86	0.56	0.55	0.62	0.60
Artificial Neural Network	0.86	0.49	0.50	0.50	0.51
Random Forest	0.14	0.49	0.50	0.50	0.49
Support Vector Machine	0.14	0.49	0.50	0.50	0.49

In terms of F1-score, LR, KNN, DT, NB, and ANN algorithms using imbalanced dataset demonstrated good performance, with precision values ranging from 0.75 to 0.80. Similarly, the RUS technique yielded poor results, while other imbalanced classification methods achieved poor to moderate F1-score value. DT achieved the highest F1-score value when using ROS (0.91). Similar with accuracy, precision and recall, ANN, RF and SVM consistently exhibit very low F1-score value and NB exhibit low F1-score value across all evaluation methods (Table 6).

Table 6
F1 Score of the obesity classification modelling

Algorithm	Imbalanced	RUS	ROS	SMOTE	ADASYN
Logistic Regression	0.80	0.56	0.58	0.72	0.71
K-Nearest Neighbour	0.80	0.50	0.77	0.78	0.78
Decision Tree	0.75	0.52	0.91	0.78	0.77
Naïve Bayes	0.80	0.56	0.55	0.59	0.57
Artificial Neural Network	0.80	0.32	0.33	0.33	0.35
Random Forest	0.03	0.32	0.33	0.33	0.32
Support Vector Machine	0.03	0.32	0.33	0.33	0.32

In terms of AUC, all models trained on imbalanced datasets as well as those utilizing ANN, RF, and SVM algorithms, produced AUC values of 0.5. However, there was a slight and minimal improvement in AUC when using the RUS and ROS methods with LR and NB algorithms. Notably, the AUC increased significantly when applying the SMOTE and ADASYN methods. The AUC values were highest for the KNN and DT algorithms when using the ROS, SMOTE, and ADASYN techniques. The most notable AUC value of 0.91 was achieved when employing the ROS method in conjunction with the DT algorithm. This indicates a substantial improvement in the model's ability to discriminate between positive and negative instances. The summary of AUC is illustrated for each imbalanced methods as depicted in Figure 5. These findings highlight the effectiveness of the ROS, SMOTE, and ADASYN methods in enhancing AUC values and improving the model's discriminative power.

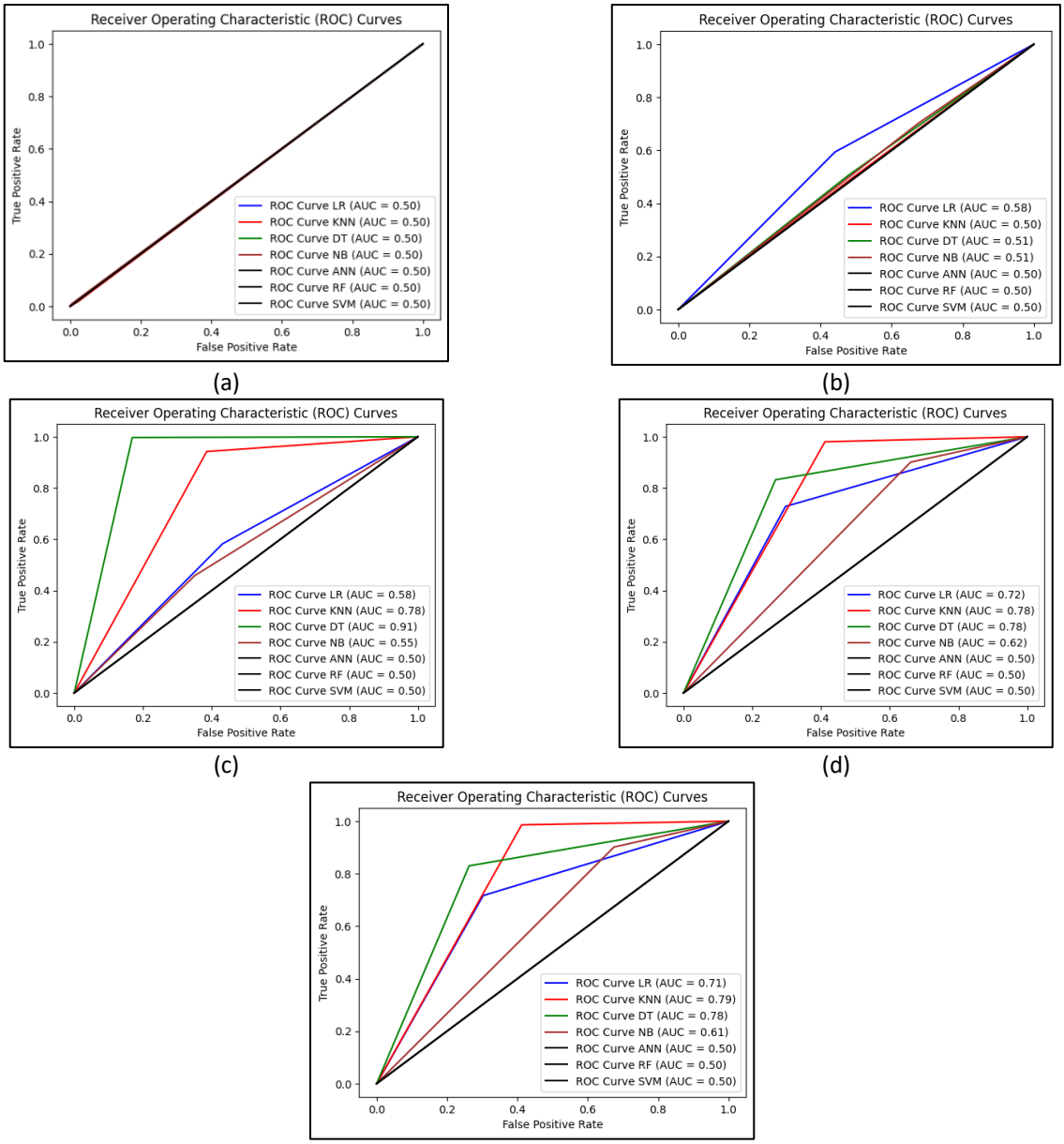

Fig. 5. The Area under the curve for ROC for each Machine Learning algorithms using (a) imbalanced dataset and multiple imbalanced techniques (b) RUS (c) ROS (d) SMOTE (e) ADASYN

5. Conclusion

Adolescent obesity has become a significant public health concern, necessitating accurate prediction and classification methods for early intervention. However, the imbalanced nature of the dataset poses challenges for developing robust classification models. This study indicates that the implementation of ROS, SMOTE, and ADASYN on the K-Nearest Neighbors (KNN) and DT algorithms consistently produced the highest accuracy, precision, recall, F1 score, and AUC for the prediction of adolescent obesity. The highest performance using all matrix was obtained using the ROS method with the DT algorithm. However, the findings also revealed that the Artificial Neural Network (ANN), Random Forest (RF), and Support vector machine (SVM) models did not exhibit satisfactory performance in predicting adolescent obesity. Despite their potential as powerful machine learning

algorithms, these models were unable to accurately classify or predict the occurrence of obesity among the target population. It is important to acknowledge that the poor performance of these models in this specific study does not imply to their ineffectiveness in other contexts or datasets. Machine learning models' performance can vary based on the nature of the data, the problem being addressed, and the specific implementation details.

5.1 Limitations and Recommendation for Future Work

This study has several limitations that should be considered when interpreting the results. Firstly, most of the data used in the study were categorical variables. This means that the analysis and modelling techniques applied may be more suitable for categorical data, potentially limiting the generalizability of the findings to datasets with predominantly numerical or continuous variables. Future studies could consider incorporating more diverse types of data to broaden the scope of analysis. Second, the feature selection process was primarily based on previous literature review and prior knowledge. Features were selected based on their relevance and consistent usage in previous studies without exploring alternative feature subsets during the modelling process. This approach may introduce biases and restrict the consideration of potentially informative features that were not previously identified. Exploring different feature subsets and employing automated feature selection techniques could enhance the robustness and comprehensiveness of the analysis.

And lastly, the hyperparameter tuning which affects the performance and generalization ability of machine learning models, was conducted using default settings only. While default settings often provide reasonable performance, they may not be optimal for every dataset or problem. Fine-tuning the hyperparameters through techniques like grid search, random search, or Bayesian optimization could potentially improve the model's performance and enhance its predictive accuracy. These limitations do not invalidate the study's findings but rather provide insights into areas where further research and improvement could enhance the analysis and modelling process. Addressing these limitations in future studies can help strengthen the validity and applicability of the findings.

Acknowledgement
This research was not funded by any grant.

References

[1] Neeharika, Chitta Hrudaya, and Yeklur Mohammed Riyazuddin. "Developing an Artificial Intelligence Based Model for Autism Spectrum Disorder Detection in Children." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 32, no. 1 (2023): 57-72. https://doi.org/10.37934/araset.32.1.5772

[2] Mohammed, Sahar Yousef. "Enhancing COVID-19 Patients Detection using Deep Transfer Learning Technique Through X-Ray Chest Images." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 32, no. 1 (2023): 290-302. https://doi.org/10.37934/araset.32.1.290302

[3] Kumar, Seema, and Aaron S. Kelly. "Review of childhood obesity: from epidemiology, etiology, and comorbidities to clinical assessment and treatment." In *Mayo Clinic Proceedings*, vol. 92, no. 2, pp. 251-265. Elsevier, 2017. https://doi.org/10.1016/j.mayocp.2016.09.017

[4] Anderson, Yvonne C., Lisa E. Wynter, Katharine F. Treves, Cameron C. Grant, Joanna M. Stewart, Tami L. Cave, Trecia A. Wouldes, José GB Derraik, Wayne S. Cutfield, and Paul L. Hofman. "Assessment of health-related quality of life and psychological well-being of children and adolescents with obesity enrolled in a New Zealand community-based intervention programme: an observational study." *BMJ open* 7, no. 8 (2017): e015776. https://doi.org/10.1136/bmjopen-2016-015776

[5] Abarca-Gómez, Leandra, Ziad A. Abdeen, Zargar Abdul Hamid, Niveen M. Abu-Rmeileh, Benjamin Acosta-Cazares, Cecilia Acuin, Robert J. Adams *et al.,* "Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128· 9 million children, adolescents, and adults." *The lancet* 390, no. 10113 (2017): 2627-2642.

[6] Pandita, Aakash, Deepak Sharma, Dharti Pandita, Smita Pawar, Mir Tariq, and Avinash Kaul. "Childhood obesity: prevention is better than cure." *Diabetes, metabolic syndrome and obesity: targets and therapy* (2016): 83-89. https://doi.org/10.2147/DMSO.S90783

[7] Mazidi, Mohsen, Maciej Banach, Andre Pascal Kengne, and Lipid and Blood Pressure Meta-analysis Collaboration Group. "Prevalence of childhood and adolescent overweight and obesity in Asian countries: a systematic review and meta-analysis." *Archives of medical science* 14, no. 6 (2018): 1185-1203. https://doi.org/10.5114/aoms.2018.79001

[8] Bibiloni, Maria del Mar, Antoni Pons, and Josep A. Tur. "Prevalence of overweight and obesity in adolescents: a systematic review." *International Scholarly Research Notices* 2013 (2013). https://doi.org/10.1155/2013/392747

[9] Institute for Public Health. "National Health and Morbidity Survey (NHMS) 2019: Vol. I: NCDs–Non-Communicable Diseases: Risk Factors and Other Health Problems." (2020).

[10] Kaur, Harsurinder, Husanbir Singh Pannu, and Avleen Kaur Malhi. "A systematic review on imbalanced data challenges in machine learning: Applications and solutions." *ACM Computing Surveys (CSUR)* 52, no. 4 (2019): 1-36. https://doi.org/10.1145/3343440

[11] Singh, Balbir, and Hissam Tawfik. "Machine learning approach for the early prediction of the risk of overweight and obesity in young people." In *Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part IV 20*, pp. 523-535. Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-50423-6_39

[12] Pranto, Badiuzzaman, Sk Maliha Mehnaz, Sifat Momen, and Syed Maruful Huq. "Prediction of diabetes using cost sensitive learning and oversampling techniques on Bangladeshi and Indian female patients." In *2020 5th international Conference on information technology research (ICITR)*, pp. 1-6. IEEE, 2020. https://doi.org/10.1109/ICITR51448.2020.9310892

[13] Wong, Jyh Eiin, Miwa Yamaguchi, Nobuo Nishi, Michihiro Araki, and Lei Hum Wee. "Predicting Overweight and Obesity Status Among Malaysian Working Adults With Machine Learning or Logistic Regression: Retrospective Comparison Study." *JMIR Formative Research* 6, no. 12 (2022): e40404. https://doi.org/10.2196/40404

[14] Devi, K. Nirmala, N. Krishnamoorthy, P. Jayanthi, S. Karthi, T. Karthik, and K. Kiranbharath. "Machine Learning Based Adult Obesity Prediction." In *2022 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1-5. IEEE, 2022. https://doi.org/10.1109/ICCCI54379.2022.9740995

[15] Ferdowsy, Faria, Kazi Samsul Alam Rahi, Md Ismail Jabiullah, and Md Tarek Habib. "A machine learning approach for obesity risk prediction." *Current Research in Behavioral Sciences* 2 (2021): 100053. https://doi.org/10.1016/j.crbeha.2021.100053

[16] Leo, Judith, Edith Luhanga, and Kisangiri Michael. "Machine learning model for imbalanced cholera dataset in Tanzania." *The Scientific World Journal* 2019 (2019). https://doi.org/10.1155/2019/9397578

[17] Kaisar, Shahriar, and Abdullahi Chowdhury. "Integrating oversampling and ensemble-based machine learning techniques for an imbalanced dataset in dyslexia screening tests." *ICT Express* 8, no. 4 (2022): 563-568. https://doi.org/10.1016/j.icte.2022.02.011

[18] Aqsha, M., S. A. Thamrin, and Armin Lawi. "Combination of ADASYN-N and Random Forest in Predicting of Obesity Status in Indonesia: A Case Study of Indonesian Basic Health Research 2013." In *Journal of Physics: Conference Series*, vol. 2123, no. 1, p. 012039. IOP Publishing, 2021. https://doi.org/10.1088/1742-6596/2123/1/012039

[19] Waqar, Muhammad, Hassan Dawood, Hussain Dawood, Nadeem Majeed, Ameen Banjar, and Riad Alharbey. "An efficient SMOTE-based deep learning model for heart attack prediction." *Scientific Programming* 2021 (2021): 1-12. https://doi.org/10.1155/2021/6621622

[20] Kokkotis, Christos, Georgios Giarmatzis, Erasmia Giannakou, Serafeim Moustakidis, Themistoklis Tsatalas, Dimitrios Tsiptsios, Konstantinos Vadikolias, and Nikolaos Aggelousis. "An explainable machine learning pipeline for stroke prediction on imbalanced data." *Diagnostics* 12, no. 10 (2022): 2392. https://doi.org/10.3390/diagnostics12102392

[21] Varotto, Giulia, Gianluca Susi, Laura Tassi, Francesca Gozzo, Silvana Franceschetti, and Ferruccio Panzica. "Comparison of resampling techniques for imbalanced datasets in machine learning: application to epileptogenic zone localization from interictal intracranial EEG recordings in patients with focal epilepsy." *Frontiers in Neuroinformatics* 15 (2021): 715421. https://doi.org/10.3389/fninf.2021.715421

[22] Oskouei, Rozita Jamili, and Bahram Sadeghi Bigham. "Over-sampling via under-sampling in strongly imbalanced data." *International Journal of Advanced Intelligence Paradigms* 9, no. 1 (2017): 58-66. https://doi.org/10.1504/IJAIP.2017.081179

[23] Khafaga, Doaa Sami, Amal H. Alharbi, Israa Mohamed, and Khalid M. Hosny. "An Integrated Classification and Association Rule Technique for Early-Stage Diabetes Risk Prediction." In *Healthcare*, vol. 10, no. 10, p. 2070. MDPI, 2022. https://doi.org/10.3390/healthcare10102070

[24] Singh, Balbir, and Hissam Tawfik. "A machine learning approach for predicting weight gain risks in young adults." In *2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, pp. 231-234. IEEE, 2019. https://doi.org/10.1109/DESSERT.2019.8770016

[25] Hossain, Rifat, SM Hasan Mahmud, Md Altab Hossin, Sheak Rashed Haider Noori, and Hosney Jahan. "PRMT: predicting risk factor of obesity among middle-aged people using data mining techniques." *Procedia computer science* 132 (2018): 1068-1076. https://doi.org/10.1016/j.procs.2018.05.022

[26] Thamrin, Sri Astuti, Dian Sidik Arsyad, Hedi Kuswanto, Armin Lawi, and Sudirman Nasir. "Predicting obesity in adults using machine learning techniques: an analysis of Indonesian basic health research 2018." *Frontiers in nutrition* 8 (2021): 669155. https://doi.org/10.3389/fnut.2021.669155

[27] Abdullah, Fadzli Syed, Nor Saidah Abd Manan, Aryati Ahmad, Sharifah Wajihah Wafa, Mohd Razif Shahril, Nurzaime Zulaily, Rahmah Mohd Amin, and Amran Ahmed. "Data mining techniques for classification of childhood obesity among year 6 school children." In *Recent Advances on Soft Computing and Data Mining: The Second International Conference on Soft Computing and Data Mining (SCDM-2016), Bandung, Indonesia, August 18-20, 2016 Proceedings Second*, pp. 465-474. Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-51281-5_47

[28] Dugan, Tamara M., S. Mukhopadhyay, Aaron Carroll, and Stephen Downs. "Machine learning techniques for prediction of early childhood obesity." *Applied clinical informatics* 6, no. 03 (2015): 506-520. https://doi.org/10.4338/ACI-2015-03-RA-0036

[29] Pang, Xueqin, Christopher B. Forrest, Félice Lê-Scherban, and Aaron J. Masino. "Prediction of early childhood obesity with machine learning and electronic health record data." *International journal of medical informatics* 150 (2021): 104454. https://doi.org/10.1016/j.ijmedinf.2021.104454

[30] Colmenarejo, Gonzalo. "Machine learning models to predict childhood and adolescent obesity: a review." *Nutrients* 12, no. 8 (2020): 2466. https://doi.org/10.3390/nu12082466

[31] Rodriguez-Almeida, Antonio J., Himar Fabelo, Samuel Ortega, Alejandro Deniz, Francisco J. Balea-Fernandez, Eduardo Quevedo, Cristina Soguero-Ruiz, Ana M. Wägner, and Gustavo M. Callico. "Synthetic patient data generation and evaluation in disease prediction using small and imbalanced datasets." *IEEE Journal of Biomedical and Health Informatics* (2022). https://doi.org/10.1109/JBHI.2022.3196697

[32] Zare, Samane, Michael R. Thomsen, Rodolfo M. Nayga Jr, and Anthony Goudie. "Use of machine learning to determine the information value of a BMI screening program." *American journal of preventive medicine* 60, no. 3 (2021): 425-433. https://doi.org/10.1016/j.amepre.2020.10.016