Early Obesity Risk Prediction via Non-dietary Lifestyle Factors using Machine Learning Approaches

Ker Ming Seaw,[1] Melvin Khee Shing Leow,[1,2,3,4,5], Xinyan Bi,[1]*

[1] Clinical Nutrition Research Centre (CNRC)
Singapore Institute of Food and Biotechnology Innovation (SIFBI)
Agency for Science, Technology and Research (A*STAR)
14 Medical Drive #07-02, MD 6 Building Yong Loo Lin School of Medicine, Singapore 117599, Singapore

[2]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

[3]Department of Endocrinology, Tan Tock Seng Hospital, Singapore

[4]Human Potential Translational Research Programme (HPTRP), Yong Loo Lin School of Medicine, National University of Singapore (NUS)

[5]Cardiovascular and Metabolic Program, Duke-NUS Medical School, Singapore

*Correspondence: bi_xinyan@sifbi.a-star.edu.sg

Keywords: *Obesity, Prediction, Non-dietary lifestyle factors, Machine learning, Artificial intelligence*

What is already known on this topic

- Beyond dietary factors, non-dietary lifestyle factors also greatly influence obesity.
- Existing studies have shown and proved that machine learning models can predict obesity, but they often overlook the complexities of non-dietary determinants, which can be crucial for early identification and prevention.

What this study adds

- This study highlights the potential of non-dietary lifestyle factors, alongside family history and demographics, as significant predictors of obesity.
- It demonstrates that machine learning models can effectively predict obesity based on these non-dietary factors, which can be vital to early identify individuals at risk of obesity to prevent or provide timely intervention.
- Results from this study provide a solid foundation for future research aimed at analyzing non-dietary factors and refining the predictive models for obesity prediction across different populations.

Abstract

Obesity poses a significant health threat, contributing to the development of noncommunicable diseases (NCDs). Early identification of individuals at higher risk for obesity is crucial for implementing effective prevention strategies. This study explores the viability of non-dietary factors such as lifestyle, family history, and demographics as predictors of obesity risk. The dataset comprised 1068 males and 1043 females, aged between 14 to 61 years. Only non-dietary factors were used to build the machine learning models, including decision tree, random forest, support vector classification (SVC), k-nearest neighbor (KNN), and Gaussian Naïve Bayes (GNB). Random forest emerged as the optimal model, demonstrating 66.9% test accuracy, 66.4% precision, 66.9% recall, 66.4% F1-score, 94.5% specificity and 92.3% area under the receiver operating characteristic curve (AUC-ROC). Variability of the models' performance was also evaluated through bootstrapping. Lifestyle factors, while less impactful than family history and demographics, also contributed to predictive power. This indicates the potential for predicting obesity while relying less on dietary data, paving the way for future studies to refine predictive models. This could play a crucial role in identifying lifestyle factors as predictors of obesity, thereby preventing and intervening early to address obesity-related complications.

1 INTRODUCTION

1.1 Background

Obesity manifests when an individual experiences an excessive increase in body weight beyond a specified threshold that may be detrimental to health status [1]. The World Health Organization (WHO) has adopted the body mass index (BMI) as a key indicator for classifying weight status, officially defining obesity as a condition characterized by a BMI of 30.0 or higher [1]. Despite concerted efforts to combat obesity, it continues to exhibit an upward trajectory. According to projections by the World Obesity Federation (WOF), approximately one billion individuals globally will be affected by obesity, comprising approximately one in seven men and five women in 2030 [2]. This projection equates to roughly one in twenty men and eleven women in Southeast Asia [3].

Individually, obesity engenders a myriad of health complications. Current clinical evidence underscores the association between obesity and an extensive spectrum of diseases, including diabetes, dyslipidemia, hypertension, cardiovascular disease (CVD), chronic kidney disease (CKD), non-alcoholic fatty liver disease (NAFLD), respiratory disorders, gastroesophageal reflux disease, and certain forms of cancer [4-6]. Even in the absence of these complications, individuals affected by obesity remain at high risk for future cardiovascular-related complications and elevated all-cause mortality rates longitudinally [7-10]. Therefore, it becomes paramount to proactively prevent the onset of obesity. To effectively mitigate the rising prevalence of obesity, the emphasis must shift from treatment to prevention. As such, the identification of individuals at high risk of developing obesity becomes important.

Machine learning is a technique of artificial intelligence in analyzing data based on statistics and probability for building models to make predictions of an outcome [11]. There is a significant rise in the application of machine learning in healthcare sectors for disease prediction in recent years [12]. Promising results have been obtained to predict disease such as diabetes and CVD using machine learning [13-14]. Previous systematic reviews have revealed that in most of the existing obesity prediction studies, a broad range and combination of factors were used as attributes [15-16]. These include data related to diet, physical activity, socio-demographics, hospitalization, and electronic medical records. Among these factors, diet and physical activity were understandably utilised most frequently. Previous meta-analysis has found that the lack of physical activity and an increase in sedentary behaviour were associated with obesity [17].

Besides exercise, non-exercise activity thermogenesis (NEAT) also contributes to the total daily energy expenditure (TDEE) [18]. Therefore, more and more studies examined the roles of other lifestyle factors, including

screen time [19-20] and the distance and accessibility to public transportation [21-22], in maintaining weight and preventing obesity. Additionally, smoking was found to be associated with obesity [23]. These results, when combined, suggested that non-dietary lifestyle factors play a significant role within the intertwined nature of obesity-related factors, which warrants the need to identify individuals susceptible to obesity based on their non-dietary lifestyle influences. This is especially true given studies that have determined that physical activity level, smoking status, and time spent watching television are independently associated with weight change [24]. With TDEE, it suggests that weight gain can still occur even if an individual is eating healthily but the energy expenditure is less than the energy intake [25]. This strongly suggest that non-dietary lifestyle factors play an integral part in weight maintenance.

Dietary information often comes with limitations [26]. Assessing dietary intake can be challenging due to the difficulty in measuring it accurately as no single method can provide a perfect assessment of dietary exposure. While methods such as 24-hour dietary recall (24HR), dietary records (DR), dietary history, and food frequency questionnaires (FFQs) are commonly used, they are subjective methods and may have limitations due to their reliance on self-reporting. This may cause inaccuracies in the information gathered resulting from inherent bias, human errors, recall issues, and methodological domain [26].

The accuracy of machine learning models can be considerably affected especially if there are several errors and inaccuracies from dietary information. It can also be implied from current evidence that the possibility of non-dietary predictors of obesity, such as physical activity levels, manifesting before significant changes in dietary habits cannot be eliminated.

With the abovementioned and considering the limitations of the various FFQs together with the sole impact from lifestyle and demographic factors on obesity, it is worth to evaluate the feasibility of predicting obesity using machine learning approach with only lifestyle factors that has also considered demographics information and compare its performance with other machine learning studies, thereby also eliminating errors and biases of dietary intake from FFQs.

However, the utilization of exclusively non-dietary lifestyle factors as predictive variables for obesity remains relatively underexplored within the scientific domain. Consequently, the outcomes derived from machine learning techniques exclusively reliant on such factors remain largely undetermined and inadequately substantiated.

1.2 Literature Review on Related Work

Notwithstanding the different methodology and different final features used, the dataset used in this study was also used by several other studies to perform machine learning within the same broad domain and topic to predict

obesity. Previously, Yagin et al. adopted a trained neural network approach and optimized with Bayesian optimization on the dataset for multiclass classification prediction [27]. However, they included both physical activity and eating habits attributes for their modelling and their mean accuracy of 10 trials with the data were high with 93.06%, accounting for every target BMI classification, from underweight to obesity type III.

On the other hand, Kaur et al. included only eating habits from this same dataset and compared between random forest, bagging classifier, gradient boost (GB) classifier, extreme gradient boost (xGB) classifier, support vector machine (SVM), and k-nearest neighbor (KNN), with various train-to-test data ratio [28]. They found that xGB and GB classifier were the overall 2 highest models at different training to testing data ratio with accuracies above 97.0%. Also using the same data source accounting for both lifestyle and dietary factors, Rodriguez et al. applied 8 models, namely, decision tree, SVM, KNN, Gaussian Naive Bayes (GNB), multilayer perceptron (MLP), random forest, GB, and xGB [29]. Random forest was the best performing model with a result of 77.69% for accuracy, 78.15% for recall, 78.53% for precision, and 78.09% for F1-score.

Beyond this dataset, several other studies had also used different datasets to predict obesity. For instance, Thamrin et al. used the data from 2013 Indonesian Basic Health Research (RISKESDAS) and accounted for demographic, physical activity, lifestyle, and dietary factors [30]. Logistic regression, Naïve Bayes, and classification and regression trees (CART) were used in their study. Across the 10 different folds, CART had better results for sensitivity (≥ 82.70%) and F1-score (≥ 72.82%), while logistic regression was the overall best model with better results for accuracy (≥ 72.20%), specificity (≥ 71.05%), precision (≥ 69.51%), kappa (≥ 44.41%), Fβ (≥ 70.30%), and AUC (≥ 79.77%).

On the other hand, Cheng et al. used the National Health and Nutrition Examination Survey (NHANES) data from 2003-2004 and 2005-2006 where their main attributes included only physical activity level and sociodemographic data, without using any dietary data [31]. Eleven models were used in their study, including MLP, random tree, classification via regression (CVR), logistic regression, naïve Bayes, Radial Basis Function (RBF), local KNN, random subspace, decision table, multiobjective evolutionary fuzzy classifier and J48. The highest AUC of 64.3% was achieved with both CVR classifier and random subspace. Random subspace showed the best result for accuracy (67.03%) and sensitivity (56.8%). Meantime, Naïve Bayes had the highest specificity (76.7%) but with lower AUC (62.0%).

1.3 Objective

In the current study, the feasibility of predicting obesity without dietary factors was explored. The aim was to apply and compare various supervised machine learning algorithms on publicly available data to determine the

best algorithm for predicting obesity from using solely non-dietary data, comprising of lifestyle, family history, and demographic data. Bootstrapping was performed to validate and better evaluate the performance variability of the models built. Models were then built with dietary data from the same data source using the same methodology, and the test accuracy between models built with non-dietary and dietary data was compared for performance.

2 METHODS

All the work in this study was performed with Python in Jupyter Notebook 6.4.8 from Anaconda Navigator 2.4.1. The high-level overview for the workflow of the study was shown in Figure S1.

2.1 Data Source

The dataset used in this study, titled "Dataset for estimation of obesity levels based on eating habits and physical condition" was obtained from the public database Kaggle [32]. This dataset was obtained through a survey administered on a web platform available for 30 days to assess obesity levels in the general populations of Peru, Colombia, and Mexico. The obtained data underwent the Synthetic Minority Over-Sampling Technique (SMOTE) by identifying the nearest neighbour of the data point of the minority classification to balance class distributions, as detailed in a published paper in the Journal Data in Brief [33]. 2111 rows by 19 columns in total, the dataset comprises of 1068 males and 1043 females, aged between 14 to 61, with a broad range of physical conditions. The data obtained with the questions asked in the web-based survey directly captures the activities happening in daily lives such as, but not limited to, time using technology devices, transportation used, and the exercise frequency. Therefore, this is very relatable and appropriate given the aim of this study and the impact of daily lives on obesity. All the questions asked with their answer selection in the survey can be interpreted by the column, "description" and "data categories" in Table S1.

2.2 Data Pre-Processing

2.2.1 Data Understanding

The data was first ensured that there was no missing and duplicated data after it was loaded using pandas. Thereafter, the dataset was checked for its data type, number of unique values in the respective columns, data value and data size including number of rows and columns, as well as what do they represent. The height in metres was then rounded off to 2 decimal places before BMI was re-calculated to ensure its accuracy. Based on the demarcation from WHO and the published paper in the Journal Data in Brief [33], the dataset was then categorized and re-assigned their classes in the column "NObeyesdad" based on their BMI (Table 1). Bar graph was then plotted to visualize the distribution of the target variable classification (Figure S2).

177 2.2.2 Exploratory Data and Pre-Modelling Analysis

178 As the aim of this research was to study between non-dietary lifestyle factors and obesity, the relationship between
179 BMI and non-dietary lifestyle attributes, i.e. - exercise frequency, main transportation used, time using technology
180 devices, and smoking status was visualised and analysed using boxplot, accounting for both genders (Figure 1).

181 2.2.3 Feature Selection

182 Considering the aim of this study was to predict outcomes using non-dietary factors, BMI and weight, along with
183 all other dietary factors, were excluded from the analysis. The remaining data was then divided into numerical
184 and categorical subsets. The numerical data, which included age and height, underwent normalization to ensure
185 the values were within a comparable range, preventing features with larger magnitudes from dominating the model
186 training process while preserving the relative differences between data points. The categorical data, which
187 comprised gender, family history of overweight, smoking status, exercise frequency, time using technology
188 devices, main transportation used and obesity status, was subjected to ordinal encoding to make it suitable for
189 machine learning models while preserving the inherent order of the data's unique values. Finally, both the
190 numerical and categorical datasets were combined into a single dataset.

191 After the dataset was combined, with the obesity status as the target variable, feature selection was performed
192 using a forward selection procedure. This method started with no features and iteratively added the feature that
193 improved the model's Akaike Information Criterion (AIC). The process continued until no further improvement
194 in the model's AIC was achieved. From the forward selection, smoking status was eliminated while the rest of the
195 features were retained as the final dataset to be used for model training (Table S2).

196 2.3 Split Data

197 2.3.1 Splitting Features and Target Variable

198 The feature matrix, consisting of the inputs, was created by dropping the target variable "NObeyesdad" which
199 reflected the obesity category, from the dataset while the target variable consisting of NObeyesdad" was extracted
200 as a separate single column for supervised learning.

201 2.3.2 Train-Test Split

202 The dataset used in the current study, consisting of only 2111 rows, was relatively small. Simple modelling
203 without complex modification or fine-tuning was performed. To ensure robust model evaluation, the final data
204 was divided into training (70%) and test (30%) sets, rather than the typical 80% training and 20% testing split.
205 This approach was taken to prevent overfitting and provide a more accurate estimation of the models' performance

on unseen data with a larger testing set. As a result, 1477 out of 2111 rows of data were used to train the model while the remaining 634 rows of data were used as test samples in getting the results for every model.

To ensure fairness in evaluating the models, the data split was performed with stratification to maintain the same distribution of obesity categories in both the training and test sets. Both datasets were then converted into Numpy arrays for compatibility with machine learning models. A fixed random state (42) was used to guarantee the reproducibility of results, ensuring consistent outcomes each time the code is executed.

2.4 Models Application, Training and Validation

2.4.1 Models Application

Five machine learning models including decision tree, random forest classifier, Support Vector Classification (SVC), GNB, and KNN, were defined and imported to be used in the current study.

The structure of decision tree resembles a hierarchical, tree-like model, starting with the most significant input as the root node. Based on conditions and subsequent selections made, this root node is split into decision nodes and ultimately, into multiple terminal nodes to give a final output [34]. Unlike decision tree which builds only one tree at a time, random forest builds multiple decision trees at a single time during the training phase from the bootstrapped dataset created by randomly selected data. The number of trees is a user-defined parameter, and the final result is determined by the most common outcome across all the trees in the forest [34].

SVC is a specific implementation within SVM used for classification purpose. The structure of SVC is relatively complicated, but it could be able to process multiple variables simultaneously. This is because SVC uses hyperplane to separate the data into 2 classes, with the numbers of hyperplanes determined by the number of independent variables. Support vectors are the points that are close to the hyperplane and final prediction is determined by these corresponding support vectors and the value of independent variables [34].

Gaussian is one of the common variants of Naïve Bayes which works by assuming independence between each pair of features [35]. Gaussian Naïve Bayes assumes a normal data distribution and adopts the approach of multiplying classification outcome probability with the likelihood of every single independent variable to result in a scoring. Thus, the output variable was predicted by every attribute independently. Consequently, the classification with the higher resulted scoring is selected as the final prediction [36].

K-nearest neighbor is a classification algorithm that works by collecting training data as instances in an n-dimensional space. The training data is categorized into clusters based on their shared characteristics, with data points within the same category being close to each other in the n-dimensional space. Test data is predicted for their outcome from their measures of similarity, that is, the shortest distance between their data points with the

training data points from various clusters. This k-value is the number of neighboring data points from the training data to include for determining the test data; this number is a hyperparameter that could be fined-tuned by the user. The final prediction would be the cluster from training data with the highest number out of all the pre-set k number of data points that are closest in terms of distance to the test data points [35, 37].

2.4.2 Model Training and Validation

Following train-test split and importing the machine learning models, a 10-fold cross-validation strategy was employed. This technique involved splitting the 70% training data into 10 equally sized folds, where each fold within the 70% training data served as a testing set once while the remaining 9 folds were used for training. The evaluation process was repeated 10 times, once for each fold, to ensure that every data point was used for both training and testing. This helped mitigate bias that could arise and provided a more reliable estimate of model performance. Stratified sampling was performed by using the StratifiedKFold method to ensure the proportion of each class in the target variable was similar across all folds.

In every model, for each iteration in the 10-fold cross-validation of the 70% training data, 9 folds were used to train the model, while the remaining 1-fold was held out as a testing set. The model was trained on the 9 training folds using the fit() method. After training, the model's performance was assessed on the test fold by generating predictions using the predict() method and predicted probabilities using the predict_proba() method. This process was repeated for all 10 iterations and folds for each model. Once the cross-validation process was completed for each model, the mean and standard deviation of each performance metric were calculated across all 10 iterations. Once the entire cross-validation was completed, the final model was trained on the full 70% training data and results were aggregated by testing the final model on the 30% testing data. Specificity was calculated using a custom function that derives specificity from the confusion matrix, while multi-class AUC-ROC was computed using the One-vs-Rest approach.

Apart from the metrics, feature importance was analyzed during the training process in every run of the 10-folds. This iterative process allows each model to be trained and evaluated on different subsets of the data, thereby ensuring each model being tested on different data points to enhance the reliability of the performance metrics. The average performance metrics were then calculated across all folds to assess the effectiveness of each model for every performance metrics. The order of performance was then visualised with bar graphs (Figure 2).

While feature importance values were directly calculated for decision tree and random forest, permutation importance was calculated for SVC and KNN. This involves shuffling each feature and measuring the decrease

in model accuracy, providing an interpretation of feature significance. No feature importance was calculated for GNB.

To estimate the variability of the performance metrics, following the 10-fold cross validation phase, a bootstrapping approach configuring with 1,000 resamples from the 70% training data was employed. 10-fold cross validation was also applied on each bootstrap sample whereby after cross-validation on each bootstrap sample, the model was evaluated on the held-out 30% testing dataset partitioned out from the original dataset. This process was repeated for 1000 iterations for the 1000 resamples, and the performance metrics across the 1000 bootstrap iterations on the same testing set was aggregated, alongside with their standard deviation and the confidence interval (95%) lower and upper values that were calculated for each metric using percentile-based methods.

2.5 Performance Analysis

Performance metrics was required to evaluate the performance of the models by using several measurements to assess their results for clearer visualization (Table 2).

Abovementioned processes for steps including feature selection, splitting data, models' application, training, validation and performance metrics evaluation (2.2.3 to 2.5) were also repeated with dietary factors from the same data source (Table S1). This is to evaluate how the model built with non-dietary data performed as compared to model built with dietary factors. These dietary factors include frequency of high calorie food consumption, frequency of vegetable consumption, frequency of food consumption between meals, frequency of alcohol consumption, number of main meals (daily), consumption of water amount (daily) and status for calorie consumption monitoring.

3 RESULTS

3.1 Pre-Modelling Analysis

From the pre-modelling analysis performed as abovementioned, several noteworthy insights were revealed. Figure 1a shows that individuals who exercised 4 to 5 times per week had a lower BMI compared to those with lower exercise frequency. It could also be observed, based on the median BMI, that except for male individuals who exercised 1 to 2 times per week, the BMI decrease as the exercise frequency increase. This observation was consistent with some of other previous studies where the results show that the increased physical activity levels potentially decreased the weight gain [41].

Unlike previous studies showing the reverse associations between the use of public transportation and BMI or body weight [42-43], results in the current study did not display any significant differences in the median BMI between public transportation and motorbike or automobile. Male individuals who used motorbikes has a lower

median BMI than those who used public transportation or automobile (Figure 1b). Additionally, the evaluation of bicycle-based transportation could not be fairly compared with other modes of transportation, primarily due to the paucity of available data pertaining to individuals who commute using bicycles, resulting in an absence of data for female individuals who commutes using bikes (Figure 1b). The reverse association between BMI or weight and the use of public transportation may be attributed to the additional walking inherent in such transportation modes. Previous study shown that public transportation was associated with additional average 8.3 more minutes of walk daily [44]. This finding aligns cohesively with the outcomes depicted (Figure 1b), which underscores that walking is correlated with a significantly lower BMI range and median when compared to other modes of transportation.

Sedentary lifestyle and screen time were found to be associated with obesity. An inclination towards a sedentary lifestyle often coincided with increased screen time. However, female individuals who spent more than 5 hours daily on technology devices has a lower median BMI (Figure 1c), whereas male individuals across all categories of time using technology devices have no significant difference in their median BMI.

As with all analysis done in this study, it was imperative to contextualize these observations within the scope of data availability. Specifically, the dataset yielded a more robust representation of individuals engaging in technology device usage for less than 5 hours daily. Notably, there were 952 individuals who use 0 to 2 hours and 915 individuals who use 3 to 5 hours of technology device daily, in contrast to the relatively limited number of 244 individuals who use more than 5 hours of technology device daily, which could have skewed the result. While females who spend 0 to 2 and more than 5 hours on technological devices per day had lower median BMI than male, females who spend 3 to 5 hours on technological devices per day had higher median BMI than male. Nonetheless, it should be noted that as an uncontrollable factor, gender has a complicated physiological relationship with obesity [45-46].

Median BMI did not appear to differ much between individuals of both genders who do not smoke, although their median BMI reflected to be overweight but not obese. Females who smoke had a significant lower BMI range and median than males who smoke (Figure 1d). Various studies have shown that unexpectedly, individuals who used to smoke are more likely to be affected by obesity than individuals who have never smoke and individuals currently smoking [47-48], whereas individuals currently smoking are more affected by a higher BMI than individuals who have never smoke [47-50]. This was the case for the males as reflected by males who smoke have significantly higher median BMI than those who do not smoke. On the contrary, results from the pre-modelling analysis showed that individuals who smoke had a lower BMI range than individuals who do not smoke.

3.2 Performances of Machine Learning Models

Table 3a and 3b summarize the results for all the machine learning models by all the above-mentioned performance evaluation metrics and their best performing model for the different parameters. The approach used to train and develop the machine learning models using the non-dietary lifestyle data was also applied on the dietary data from the same data source. The differences in test accuracy on the dietary and non-dietary models without bootstrapping were compared and visualised (Figure 3).

3.3 Feature Importance

Feature importance measures the impact of each feature to the predictive performance of the models by calculating internally during the training process. It works by calculating how much does the features contribute to reduce the impurity as the model is being trained [51], and this was directly calculated for random forest (Figure 2a) and decision tree (Figure 2b) Meanwhile, permutation importance assesses feature importance by shuffling the values of each feature and evaluate its impact on the models' performance, thereby determining which features would influence more in predicting the output [51]. In this study, this was calculated for KNN (Figure 2c) and SVC (Figure 2d). Conversely, GNB does not learn feature importance or coefficients during training since there is an assumption of feature independence, so evaluation of features was not performed with GNB in this study.

4 DISCUSSIONS

4.1 Findings

The results in table 3 were derived from the 10-fold cross-validation of all the model algorithms, with values averaged across all folds, rather than showing results for each individual fold. For the non-dietary lifestyle data, random forest classifier was the best performing model with the highest values for all metrics, having 67.8% (± 3.2) test accuracy, 67.3% (± 3.5) precision, 67.8% (± 3.2) recall, 67.1% (± 3.4) F1-score, 94.6% (± 0.5) specificity and 91.4% (± 1.1) AUC-ROC. To better evaluate the variability of the models' performance, bootstrapping was performed with the non-dietary data. Bootstrapping increased all the metrics values for all models. Consequently, there was a change in the order of performance between non-bootstrap and bootstrapped data in all the metrics except for F1-score. Nonetheless, the ranking of performance for random forest remains unchanged as the top performer, reinforcing its position as the best-performing model.

ROC is a curve that plots sensitivity against threshold of varying values, connecting different extremes, while the AUC averages the accuracy across all threshold values [52]. In this study, AUC-ROC was calculated from classification probability predictions, rather than from single discrete classifications, providing a more

comprehensive assessment of model performance. An AUC between 0.7 and 0.8 is considered acceptable, 0.8 to 0.9 is excellent, and above 0.9 is outstanding [40]. With an AUC-ROC of 92.3%, the random forest model achieves an outstanding result. This may be due to the fact that feature bagging being used in random forest could help to account for variance in the dataset [34], thereby allowing predictions to be more precise.

The difference in test accuracy between models built with dietary and non-dietary data was evaluated. As shown in Figure 3, when the models were trained using the same approach with dietary data, there was a significant change in the order of the models based on their test accuracy, with decision tree being the best-performing model. Nonetheless, it is consistent that GNB was the least performing model in both cases, and test accuracy with non-dietary data was generally higher than with dietary data.

Unlike statistical significance, which determines whether an observed outcome is due to random chance or a particular factor of interest [53], feature importance interprets the ranking of magnitude of all the features in contributing to the predictive power of the models for the prediction [54]. While the most important features vary across different models, it can be deduced that demographics expectedly remain the most important features among all the different models in predicting obesity (Figure 2). It is unanticipated and noteworthy to observe that family history with overweight is the least important feature for decision tree and KNN models, while it ranks as the second most important feature for the SVC model. Regarding lifestyle features, there is variation in the importance of features across models. However, exercise frequency and time spent using technology devices daily retain their predictive power, constantly ranking as mid-performing features across all models, except for SVC, where the mode of transportation is more important in predicting obesity.

4.2 Comparisons with Other Work

Predictably, the results of the models in this study, which were trained using non-dietary data without bootstrapping, were not better than those of studies by Yagin et al. [27], Kaur et al. [28], and Rodriguez et al. [29], all of whom used the same data source. In the study by Yagin et al., their mean accuracy was higher at 93.06% and they also predicted multi-class classifications for different weight statuses, incorporating both physical activity and eating habits in their trained neural network. In the study by Kaur et al., only eating habits data was used, with several train-to-test data ratio tested. The highest accuracy, above 97.0%, was achieved by xGB and GB models. In the study by Rodriguez et al. [29], both lifestyle and dietary factors were included. Similarly, random forest was the best-performing model in their study. Their results were higher, with 77.69% for accuracy, 78.15% for recall, 78.53% for precision, and 78.09% for F1-score.

Currently, numerous studies used dietary data to predict obesity via machine learning algorithms. However, there is always a certain degree of errors when collecting dietary data because situations such as challenges in dietary recall or oversimplification of dietary pattern could not provide adequate crucial details that are vital for data collection to build the model [55]. To the best of our knowledge, this is the first machine learning research which simultaneously accounts for uncontrollable factor, e.g. demographics and family history, and includes solely lifestyle data without any dietary data.

4.3 Limitation and Value

We acknowledge the potential limitations and challenges in this study. Firstly, the data utilized in this study was sourced from publicly available database instead of being collected through clinical studies. Secondly, there was potential to augment the dataset with a larger volume of data. While there was a lack of external validity to evaluate how the model would perform on data beyond the dataset it was trained on, the 70-to-30 train-to-test ratio, compared to the 80-to-20 split, allowed for a larger test set, which provided better generalization to unseen data. A comprehensive 10-fold cross-validation strategy on the training data was executed in this investigation, ensuring that all data points within the training dataset were ultimately utilized for both training and testing purposes. At the same time, testing the model on the 30% held-out testing data gives an additional layer of validation. This ensured that the model was not solely dependent on the training data or internal validation procedures but could also be assessed on an entirely separate portion of the dataset, thereby stimulating how the model might perform on completely unseen data.

Additionally, results presented herein constituted an aggregation of performance metrics from all the 10 folds, thus any bias obtained through misclassification estimates could be minimized. This could ensure a more robust assessment of the efficacy of the machine learning algorithms, therefore allowing the study to be more reliable.

Furthermore, while the metric values without bootstrapping was taken as the result of the models built, additional set of results were calculated from non-dietary data that was subjected to bootstrapping to better evaluate models' performance variability. While bootstrapping increased all the metrics values significantly, the order for most of the performance of the models remains the same, which reinforces the predictive ability of the best performing model – random forest. Besides, lower and upper values for 95% confidence intervals was also calculated in addition to the standard deviation. This allows a better understanding of the range of likely values for the true performance of the model. With this, it highlights the potential for training and application of such models on other dataset from other regions in the future.

The aim of this study has been proven - there exists a potential feasibility of utilizing non-dietary determinants as factors to early predict and identify the obesity risk based on lifestyle influences. While this may not accurately represent the global or other regional populations due to the limited dataset, the result from this study offers a valuable starting point for exploring the potential of predictive modelling across different regions, providing a solid foundation for future research aimed at refining the models for different populations.

It may also be disputed that obesity prediction may be needless since diagnosis is not difficult. Nonetheless, such study can provide a simple model for first-stage screening of potential cases, whereby high-performing models can be applicable to broad and large populations on a greater scale in general. With prevention of obesity as an objective, this can in turn aids in public health planning and allocation of resources from a macro perspective.

Unlike deep learning, the use of traditional supervised learning can generate feature importance in this study to better understand where lifestyle factors stand as compared to family history and demographic factors across the different models in predicting obesity. It is also imperative to note that due to the paucity of existing studies exclusively focused on solely non-dietary datasets, a comprehensive refinement, fine-tuning and exploration of specific modelling algorithms was not performed. Instead, a broad application of diverse supervised machine learning techniques was employed to yield standardized baseline results, thereby providing possible future studies in the same domain and context a reference point for result interpretation and comparison.

4.4 Direction for Future Work

The commendable performance outcomes, especially the results from bootstrapping, highlight the potential for future clinical studies to replicate these findings with a more extensive dataset. Notwithstanding the study's intended scope, it is plausible to collect more extensive and detailed data from comprehensive questionnaires and larger populations in future endeavours to enhance the training data quality, for constructing more refined machine learning models. For example, the smoking status data could encompass options for individuals such as currently smoking, previously smoked, or never smoke categories and physical activity can incorporate cardiovascular or resistant-training categories, giving a more detailed breakdown of data type. Consequently, a broader and more realistic dataset could encompass a wider variance, allowing the machine learning models to be more applicable in real-world scenarios.

5 CONCLUSIONS

In conclusion, the random forest model emerged as the most proficient overall performer, displaying superior outcomes in terms of test accuracy, specificity, precision, recall, F1-score and AUC-ROC. The path towards leveraging machine learning for obesity prediction, particularly using exclusively non-dietary factors, remains

considerably ambitious and arduous at this juncture. Obesity is largely a multifactorial condition resulting from the interaction between genes and the environment. With respect to environmental factors, diet undoubtedly is key alongside other external factors including physical activity and medications that impact on metabolism, as well as the internal milieu of the body including hormonal dysfunction that disrupts appetite regulation and weight homeostasis. This study does not seek to downplay the undeniable critical impact of dietary factors on obesity; rather, it underscores the potential innovation in forecasting obesity solely through non-dietary variables as non-dietary factors may manifest before dietary habits and weight changes. The notable accuracy levels achieved in this study's algorithms, even without exhaustive parameter fine-tuning, is comparable to those which includes dietary data and underscore the importance of non-dietary factors in relation to obesity as well as the feasibility of this pursuit. More studies in the future including more relevant and precise data can improve machine learning models in predicting obesity. Deep learning algorithms can also be explored, particularly in scenarios involving massive data volumes. Ultimately, such strategies can facilitate identification of individuals who are at risk, enabling prevention and early interventions to reduce the costly morbidities that accompany the burden of obesity.

Authors' Contributions

KMS conceptualize, developed methodology, did formal analysis and writing of original draft. XB and MKSL were responsible for supervision, writing – review and editing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We acknowledge the financial support from Singapore Institute of Food and Biotechnology Innovation, Agency for Science, Technology, and Research (A*Star), Singapore.

Funding

No external funding was received for conducting this study.

Ethics Approval

Not applicable.

Conflict of Interests

The authors declare no conflict of interest.

Data Availability

Dataset used to perform the analysis is publicly available at https://www.kaggle.com/datasets/mandysia/obesity-dataset-cleaned-and-data-sinthetic. The result analysis generated during and/or analyzed during the current study are within the paper and available from the author on reasonable request.

References

1. World Health Organization. Obesity and overweight. https://www.who.int/en/news-room/fact-sheets/detail/obesity-and-overweight. [Accessed on 21 January 2025]
2. Tham KW, Ghani RA, Cua SC, et al. Obesity in South and Southeast Asia - A new consensus on care and management. *Obes Rev*. 2023; **24**(2): e13520. doi:10.1111/obr.13520.
3. Lobstein T, Brinsden H, Neveux M. World Obesity Atlas 2022. https://www.worldobesityday.org/assets/downloads/World_Obesity_Atlas_2022_WEB.pdf. [Accessed on 21 January 2025]
4. Hamjane N, Benyahya F, Mechita MB, Nourouti NG, Barakat A. The complications of overweight and obesity according to obesity indicators (body mass index and waist circumference values) in a population of Tangier (northern Morocco): a cross-sectional study. *Diabetes Metab Syndr*. 2019; **13**(4): 2619–2624. doi:10.1016/j.dsx.2019.07.033.
5. Abdelaal M, le Roux CW, Docherty NG. Morbidity and mortality associated with obesity. *Ann Transl Med*. 2017; **5**(7): 161. doi:10.21037/atm.2017.03.107.
6. Than WH, Chan GCK, Ng JKC, Szeto CC. The role of obesity on chronic kidney disease development, progression, and cardiovascular complications. *ABST*. 2020; 2: 24–34. doi:10.1016/j.abst.2020.09.001.
7. Bray GA, Heisel WE, Afshin A, et al. The science of obesity management: An endocrine society scientific statement. *Endocr Rev*. 2018; **39**(2): 79–132. doi:10.1210/er.2017-00253.
8. Roberson LL, Aneni EC, Maziak W, et al. Beyond BMI: the "metabolically healthy obese" phenotype & its association with clinical/subclinical cardiovascular disease and all-cause mortality -- a systematic review. *BMC Public Health*. 2014; **14**: 14. doi:10.1186/1471-2458-14-14.
9. Bell JA, Hamer M, Sabia S, Singh-Manoux A, Batty GD, Kivimaki M. The natural course of healthy obesity over 20 years. *J Am Coll Cardiol*. 2015; **65**(1): 101–102. doi:10.1016/j.jacc.2014.09.077.
10. Eckel N, Meidtner K, Kalle-Uhlmann T, Stefan N, Schulze MB. Metabolically healthy obesity and cardiovascular events: A systematic review and meta-analysis. *Eur J Prev Cardiol*. 2016; **23**(9): 956–966. doi:10.1177/2047487315623884.
11. Jayatilake SMDAC, Ganegoda GU. Involvement of machine learning tools in healthcare decision making. *J Healthc Eng*. 2021;2021:6679512. doi:10.1155/2021/6679512
12. Kumari J, Kumar E, Kumar D. A structured analysis to study the role of machine learning and deep learning in the healthcare sector with big data analytics. *Arch Computat Methods Eng*. 2023;1–29. Advance online publication. doi:10.1007/s11831-023-09915-y
13. Jaiswal V, Negi A, Pal TA. Review on current advances in machine learning based diabetes prediction. *Prim Care Diabetes*. 2021;**15**(3):435–443. doi:10.1016/j.pcd.2021.02.005
14. Azmi J, Arif M, Nafis MT, Alam MA, Tanweer S, Wang G. A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. *Med Eng Phys*. 2022;105:103825. doi:10.1016/j.medengphy.2022.103825
15. Triantafyllidis A, Polychronidou E, Alexiadis A, et al. Computerized decision support and machine learning applications for the prevention and treatment of childhood obesity: a systematic review of the literature. *Artif Intell Med*. 2020;104:101844. doi:10.1016/j.artmed.2020.101844
16. Alkhalaf M, Yu P, Shen J, Deng C. A review of the application of machine learning in adult obesity studies. *Applied Computing and Intelligence*. 2022;**2**(1):32-48. doi:10.3934/aci.2022002
17. Silveira EA, Mendonca CR, Delpino FM, et al. Sedentary behavior, physical inactivity, abdominal obesity and obesity in adults and older adults: a systematic review and meta-analysis. *Clin Nutr ESPEN*. 2022;50:63–73. doi:10.1016/j.clnesp.2022.06.001
18. Chung N, Park M-Y, Kim J, et al. Non-exercise activity thermogenesis (NEAT): a component of total daily energy expenditure. *J Exerc Nutrition Biochem*. 2018;**22**(2):23–30. doi:10.20463/jenb.2018.0013
19. Kerkadi A, Sadig AH, Bawadi H, et al. The Relationship between lifestyle factors and obesity indices among adolescents in qatar. *Int J Environ Res Public Health*. 2019;**16**(22):4428. doi:10.3390/ijerph16224428
20. Al-Hazzaa HM, Abahussain NA, Al-Sobayel HI, Qahwaji DM, Musaiger AO. Lifestyle factors associated with overweight and obesity among Saudi adolescents. *BMC Public Health*. 2012;12:354. doi:10.1186/1471-2458-12-354
21. Tan SB, Dickens BL, Sevtsuk A, et al. Exploring how socioeconomic status affects neighbourhood environments' effects on obesity risks: A longitudinal study in Singapore. *Landscape and Urban Planning*. 2022;226:104450. doi:10.1016/j.landurbplan.2022.104450

22. Tan SB. Changes in neighborhood environments and the increasing socioeconomic gap in child obesity risks: Evidence from Singapore. *Health Place*. 2022;76:102860. doi:10.1016/j.healthplace.2022.102860
23. Fasting MH, Nilsen TI, Holmen TL, Vik T. Lifestyle related to blood pressure and body weight in adolescence: cross sectional data from the Young-HUNT study, Norway. *BMC Public Health*. 2008;8:111. doi:10.1186/1471-2458-8-111
24. Mozaffarian D, Hao T, Rimm EB, Willett WC, Hu FB. Changes in diet and lifestyle and long-term weight gain in women and men. *N Engl J Med*. 2011;**364**(25), 2392–2404. doi:10.1056/NEJMoa1014296
25. Melby CL, Paris HL, Sayer RD, Bell C, Hill JO. Increasing energy flux to maintain diet-induced weight loss. *Nutrients*. 2019;**11**(10), 2533. doi:10.3390/nu11102533
26. Shim JS, Oh K, Kim HC. Dietary assessment methods in epidemiologic studies. *Epidemiol Health*. 2014;36:e2014009. doi:10.4178/epih/e2014009
27. Yagin FH, Gulu M, Gormez Y, et al. Estimation of obesity levels with a trained neural network approach optimized by the bayesian technique. *Applied Sciences*. 2023;**13**(6):3875. doi:10.3390/app13063875
28. 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):458–469. doi:10.1007/s12020-022-03215-4
29. Rodríguez E, Rodríguez E, Nascimento L, da Silva A, Marins F. Machine learning techniques to predict overweight or obesity. Informatics & Data-Driven Medicine: Proceedings of the 4th International Conference on Informatics & Data-Driven Medicine 2021;3038:190-204.
30. Thamrin SA, Arsyad DS, Kuswanto H, Lawi A, Nasir S. Predicting obesity in adults using machine learning techniques: an analysis of Indonesian basic health research 2018. *Front Nutr*. 2021;8:669155. doi:10.3389/fnut.2021.669155
31. Cheng X, Lin S-Y, Liu J , et al. Does physical activity predict obesity-a machine learning and statistical method-based analysis. *Int J Environ Res Public Health*. 2021;**18**(8):3966. doi:10.3390/ijerph18083966
32. Obesity dataset cleaned and data sinthetic. Sia M. Kaggle. https://www.kaggle.com/datasets/mandysia/obesity-dataset-cleaned-and-data-sinthetic. Accessed 03 July 2023
33. Palechor FM, Manotas AH. 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. doi:10.1016/j.dib.2019.104344
34. Dhillon SK, Ganggayah MD, Sinnadurai S, Lio P, Taib NA. Theory and practice of integrating machine learning and conventional statistics in medical data analysis. *Diagnostics (Basel)*. 2022;**12**(10):2526. doi:10.3390/diagnostics12102526
35. Sarker IH. Machine learning: algorithms, real-world applications and research directions. *SN Comput Sci*. 2021;**2**(3):160. doi:10.1007/s42979-021-00592-x
36. Raizada RD, Lee YS. Smoothness without smoothing: why Gaussian naive Bayes is not naive for multi-subject searchlight studies. *PloS One*. 2013;**8**(7):e69566. doi:10.1371/journal.pone.0069566
37. An Q, Rahman S, Zhou J, Kang JJ. A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges. *Sensors (Basel).* 2023;**23**(9):4178. doi:10.3390/s23094178
38. Wang Y, Jia P, Liu L, Huang C, Liu Z. A systematic review of fuzzing based on machine learning techniques. *PloS One*. 2020;**15**(8):e0237749. doi:10.1371/journal.pone.0237749
39. Fränti P, Mariescu-Istodor R. Soft precision and recall. *Pattern Recognit Lett*. 2023;167:115–121. doi:10.1016/j.patrec.2023.02.005
40. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol*. 2010;**5**(9):1315–1316. doi:10.1097/JTO.0b013e3181ec173d
41. Jakicic JM, Powell KE, Campbell WW, et al. Physical activity and the prevention of weight gain in adults: a systematic review. *Med Sci Sports Exerc*. 2019;**51**(6):1262–1269. doi:10.1249/MSS.0000000000001938
42. She Z, King DM, Jacobson SH. Analyzing the impact of public transit usage on obesity. *Prev Med*. 2017;99:264–268. doi:10.1016/j.ypmed.2017.03.010

43. Habinger JG, Chavez JL, Matsudo SM, et al. Active transportation and obesity indicators in adults from latin America: ELANS Multi-Country Study. *Int J Environ Res Public Health*. 2020;**17**(19):6974. doi:10.3390/ijerph17196974
44. Edwards RD. Public transit, obesity, and medical costs: assessing the magnitudes. *Prev Med*. 2008;**46**(1):14–21. doi:10.1016/j.ypmed.2007.10.004
45. Cooper AJ, Gupta SR, Moustafa AF, Chao AM. Sex/Gender differences in obesity prevalence, comorbidities, and treatment. *Curr Obes Rep*. 2021;**10**(4):458–466. doi:10.1007/s13679-021-00453-x
46. Kapoor N, Arora S, Kalra S. Gender disparities in people living with obesity - an unchartered territory. *J Midlife Health*. 2021;**12**(2):103–107. doi:10.4103/jmh.jmh_48_21
47. Dare S, Mackay DF, Pell JP. Relationship between smoking and obesity: a cross-sectional study of 499,504 middle-aged adults in the UK general population. *PloS One*. 2015;**10**(4):e0123579. doi:10.1371/journal.pone.0123579
48. Ginawi IA, Bashir AI, Alreshidi YQ, et al. Association between obesity and cigarette smoking: A community-based study. *J Endocrinol Metab*. 2016;**6**(5):149–153. doi:10.14740/jem378e
49. Sun M, Jiang Y, Sun C, et al. The associations between smoking and obesity in northeast China: a quantile regression analysis. *Sci Rep*. 2019;**9**(1):3732. doi:10.1038/s41598-019-39425-6
50. Mackay DF, Gray L, Pell JP. Impact of smoking and smoking cessation on overweight and obesity: Scotland-wide, cross-sectional study on 40,036 participants. *BMC Public Health*. 2013;13:348. doi:10.1186/1471-2458-13-348
51. Cava W, Bauer C, Moore JH, Pendergrass SA. Interpretation of machine learning predictions for patient outcomes in electronic health records. *AMIA Annu Symp Proc.* 2020;2019:572-581
52. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*. 2007;**115**(5):654–657. doi:10.1161/CIRCULATIONAHA.105.594929
53. Andrade C. The P value and statistical significance: misunderstandings, explanations, challenges, and alternatives. *Indian J Psychol Med*. 2019;**41**(3):210–215. doi:10.4103/IJPSYM.IJPSYM_193_19
54. Saarela M, Jauhiainen S. Comparison of feature importance measures as explanations for classification models. *SN Appl Sci*. 2021;**3**(2). doi:10.1007/s42452-021-04148-9
55. Morgenstern JD, Rosella LC, Costa AP, de Souza RJ, Anderson LN. Perspective: big data and machine learning could help advance nutritional epidemiology. *Adv Nutr*. 2021;**12**(3):621–631. doi:10.1093/advances/nmaa183

Abbreviations

WHO: World Health Organization

BMI: Body Mass Index

WOF: World Obesity Federation

CVD: Cardiovascular Disease

CKD: Chronic Kidney Disease

NAFLD: Non-Alcoholic Fatty Liver Disease

NEAT: Non-Exercise Activity Thermogenesis

TDEE: Total Daily Energy Expenditure

DR: Dietary Records

FFQ: Food Frequency Questionnaire

GB: Gradient Boost classifier

xGB: Extreme Gradient Boost classifier

SVM: Support Vector Machine

KNN: K-Nearest Neighbor

GNB: Gaussian Naive Bayes

MLP: Multilayer Perceptron

CART: Classification and Regression Trees

NHANES: National Health and Nutrition Examination Survey

CVR: Classification via Regression

RBF: Radial Basis Function

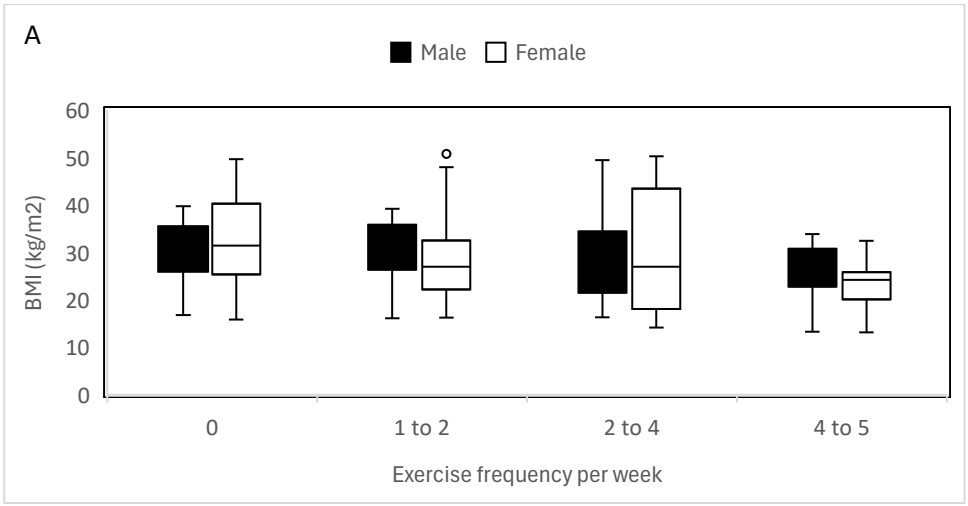
SMOTE: Synthetic Minority Over-Sampling Technique

SVC: Support Vector Classification

AUC: Area under the curve

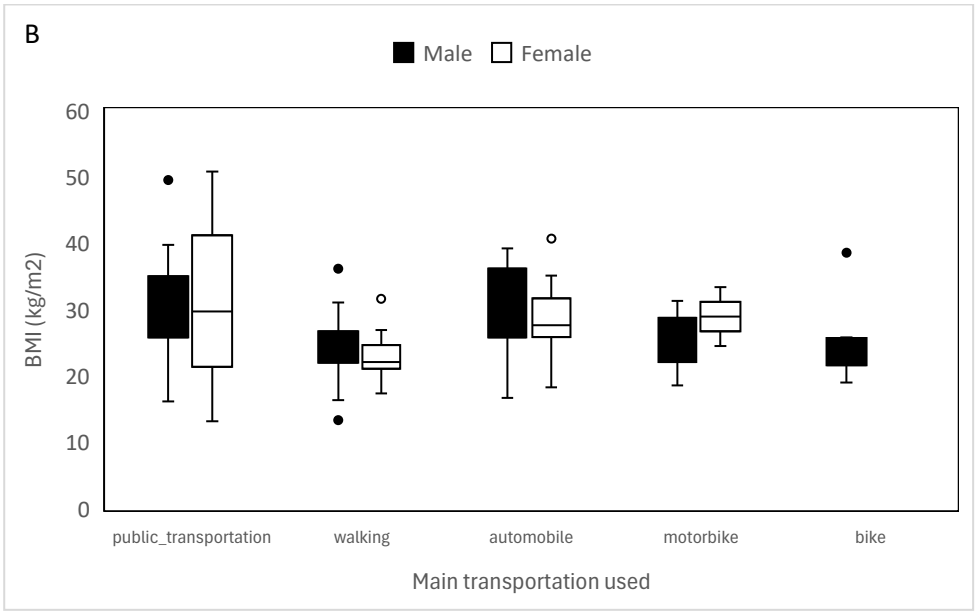
ROC: Receiver Operating Characteristic

AIC: Akaike Information Criterion

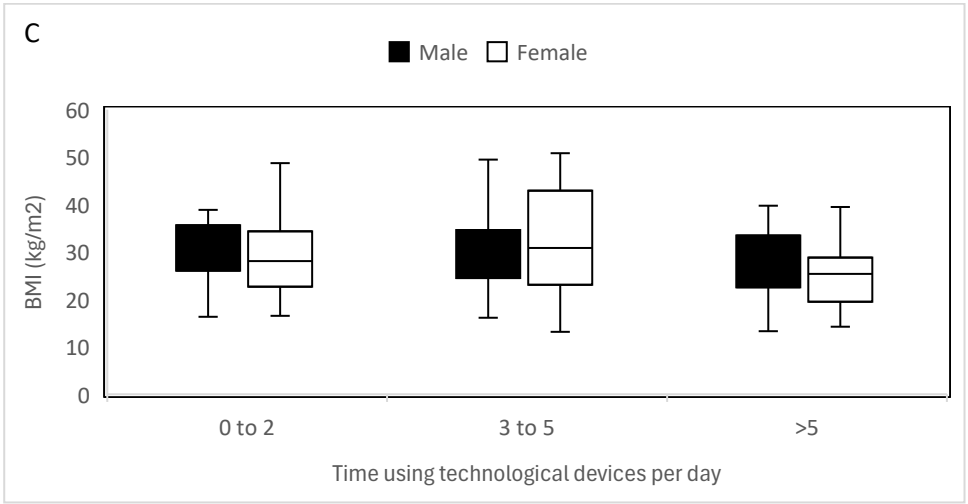


657

658



659



660

661

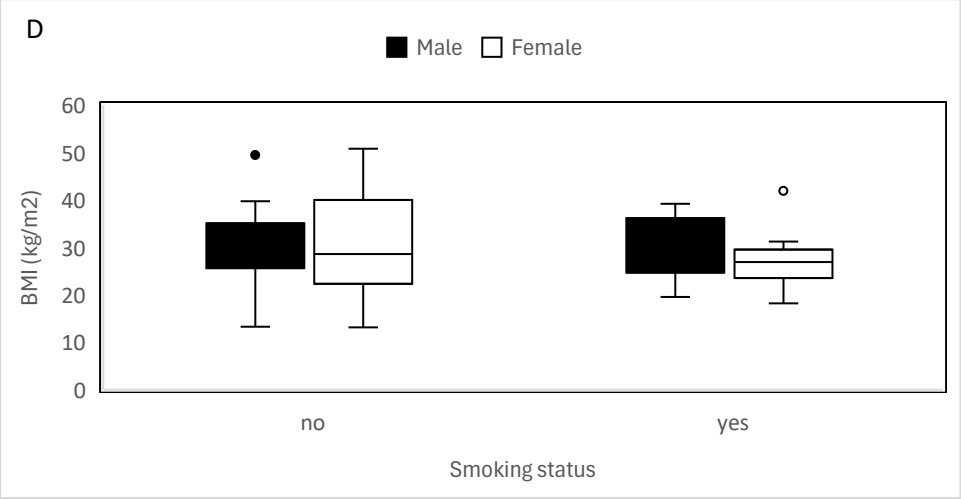
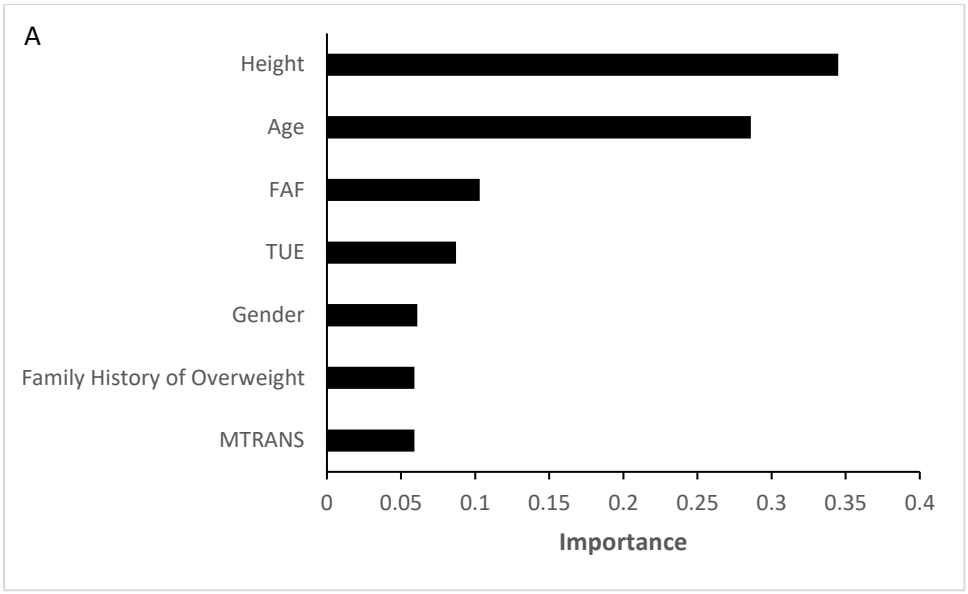
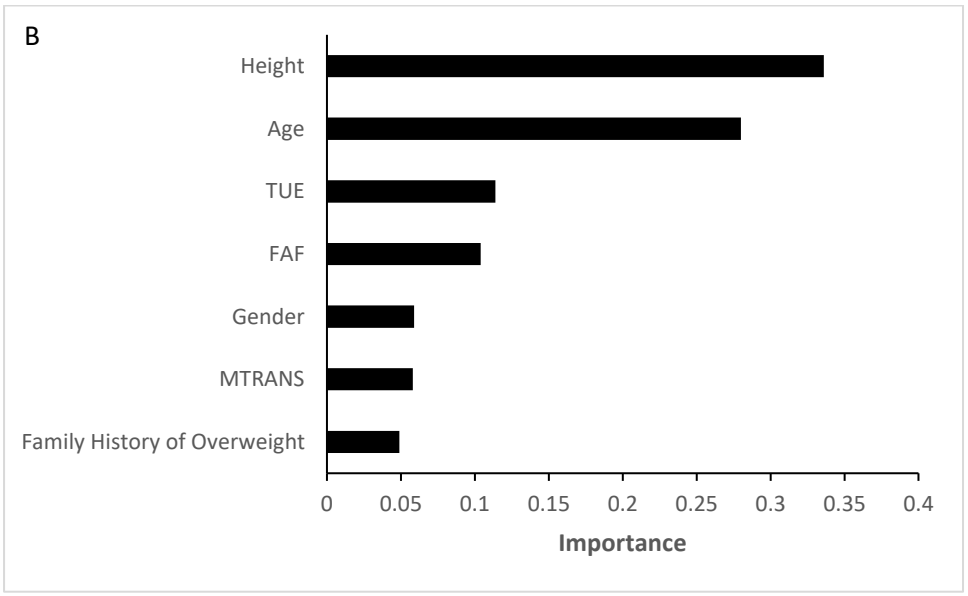


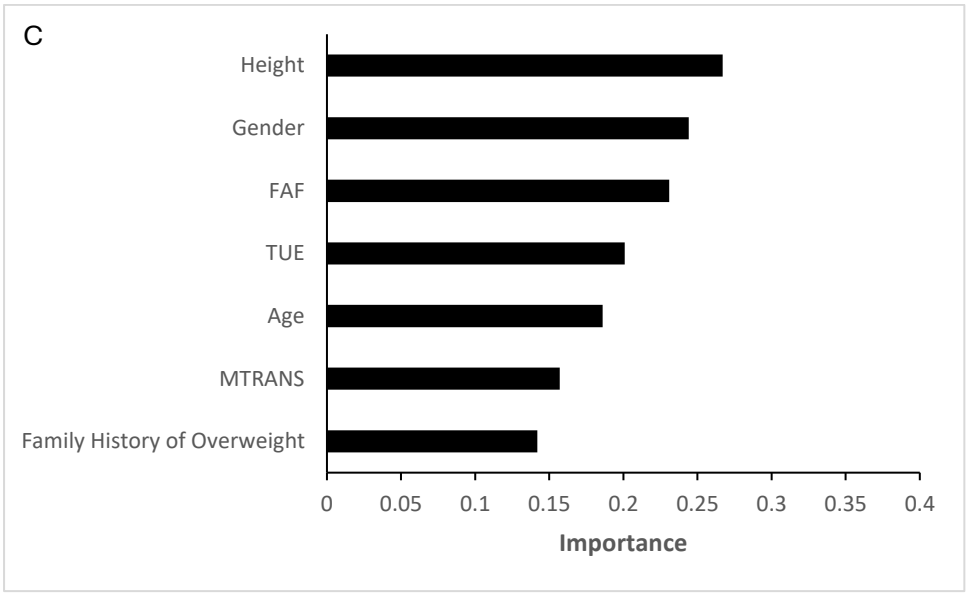
Figure 1. Box plot displaying the BMI distribution in 2111 rows of data from individuals in by (A) exercise frequency per week, (B) main transportation used, (C) time using technology devices per day, and (D) smoking status. BMI, body mass index.


A
Height
Age
FAF
TUE
Gender
Family History of Overweight
MTRANS
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
Importance

667


B
Height
Age
TUE
FAF
Gender
MTRANS
Family History of Overweight
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
Importance

668


C
Height
Gender
FAF
TUE
Age
MTRANS
Family History of Overweight
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
Importance

669

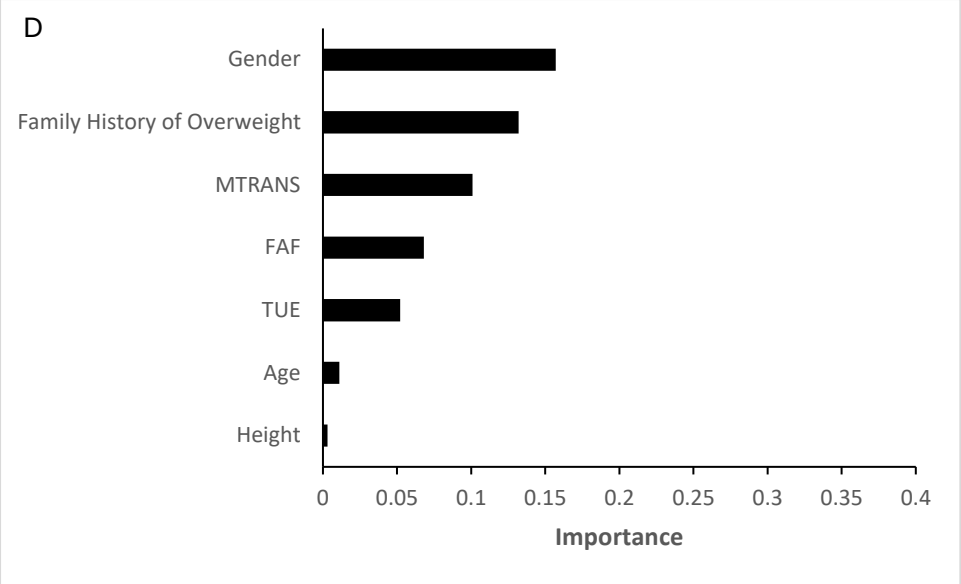


Figure 2. Feature importance of all variables for (A) random forest, and (B) decision tree. Permutation importance of all variables for (C) KNN, and (D) SVC. The importance of the different non-dietary features in terms of being predictors for obesity when being fitted into machine learning models, in comparison of all features for all the different models except for GNB, and their ranking of importance is in descending order across the various models.

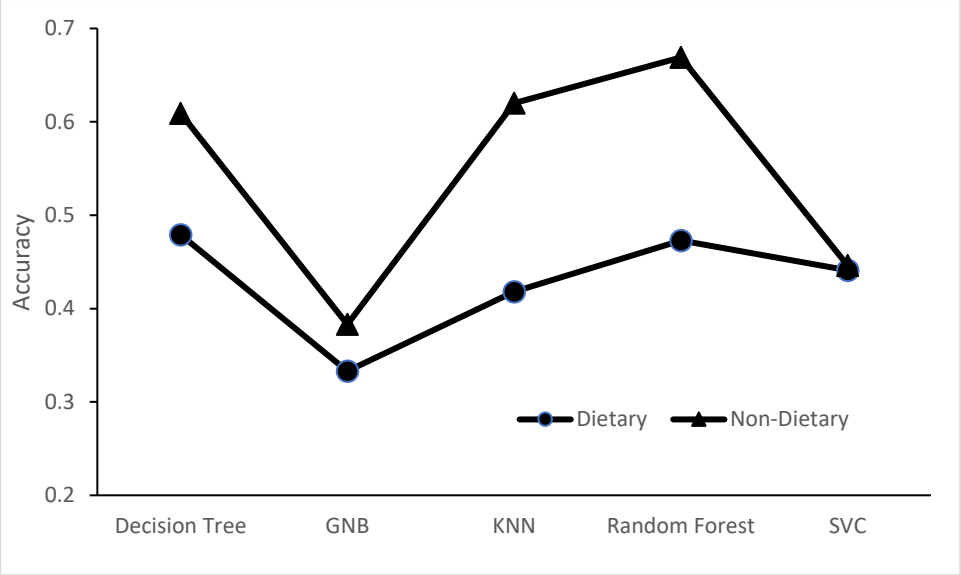


Figure 3. Test accuracy comparison of the models between dietary and non-dietary data. The graphs represent the test accuracy result of the models built from dietary and non-dietary data, respectively, without bootstrapping.

685 **Table 1.** Classification and the BMI values (N=2111).

NObeyesdad	BMI Value (kg/m^2)	Number of Rows	Proportion within dataset (%)
Underweight	Less than 18.5	267	12.6
Normal Weight	18.5-24.9	306	14.5
Overweight I	25.0-27.4	335	15.9
Overweight II	27.5-29.9	230	10.9
Obese Type I	30.0-34.9	368	17.4
Obese Type II	35.0-39.9	338	16.0
Obese Type III	40.0 and onwards	267	12.6

686 *Note:* NObeyesdad was the name as provided from the original data source, and it represents the different
687 classification for the weight status as the output variable
688 Abbreviations: BMI, body mass index
689

690 **Table 2.** Performance metrics used to assess the models.

Metric used	Purpose
Test accuracy	Proportion out of all the model's overall predictions that are correct [38].
Precision	Proportion of the model's true positive predictions among all that were predicted to be positives including false positive [39].
Recall (sensitivity)	Proportion of the model's true positive predictions among all the actual positives including both true positive and false negative [39].
Specificity	Proportion of the model's true negative predictions among all the actual negatives including both true negative and false positive [39].
F1-Score	A single metric to evaluate classification performance model by calculating the harmonic mean between precision and recall, achieving a balance between both [39].
Area under the curve (AUC)	A number to represent the performance of the Receiver Operating Characteristic (ROC) curve, in which the threshold of ROC is directly but inversely proportional to the precision and recall, respectively [38, 40]. Values for AUC range between 0 to 1 whereby a value of 0.5 indicates inability to differentiate and the higher the value above 0.5, the more superior is the performance of the models [40].

691 *Note:* All the metrics that were used to assess the machine learning models' performance in this study and their
692 interpretation
693

694 Table 3a. Overall results of the models built with non-dietary data without bootstrapping on 30% test set.

Models	Accuracy	Specificity	Precision	Recall	F1-Score	AUC-ROC
Decision tree	60.9	93.5	61.7	60.9	61.1	78.5
Random forest	66.9	94.5	66.4	66.9	66.4	92.3
SVC	44.6	90.7	40.0	44.6	39.2	81.2
GNB	38.3	89.7	42.1	38.3	29.6	74.9
KNN	62.0	93.6	60.4	62.0	60.9	86.1
Best model	Random forest	Random forest	Random forest	Random forest	Random forest	Random forest

695 *Note:* Results are reflective of the model built on the data from the data source before the application of
696 bootstrapping, and is aggregated from the 10-fold iterations.
697 Abbreviations: AUC-ROC, area under the receiver operating characteristic curve; GNB, Gaussian Naïve Bayes;
698 KNN, k-nearest neighbour; SVC, support vector classification.
699

700 Table 3b. Overall results of the models built with bootstrapped non-dietary data on 30% test set.

Models	Accuracy (95% CI)	Specificity (95% CI)	Precision (95% CI)	Recall (95% CI)	F1-Score (95% CI)	AUC-ROC (95% CI)
Decision tree	82.6±1.4 (79.8,85.2)	97.1±0.2 (96.6,97.5)	82.8±1.4 (79.9,85.3)	82.6±1.4 (79.8,85.2)	82.5±1.4 (79.8,85.1)	90.2±0.8 (88.6,91.8)

Random forest	84.4±1.2 (82.0,86.9)	97.4±0.2 (97.0,97.8)	84.5±1.3 (82.0,87.1)	84.4±1.2 (82.0,86.9)	84.3±1.3 (81.8,86.9)	96.8±0.4 (96.1,97.6)
SVC	44.4±1.5 (41.3,47.3)	90.6±0.2 (90.1,91.1)	45.3±4.0 (38.9,54.3)	44.4±1.5 (41.3,47.3)	39.8±1.8 (36.1,43.2)	80.9±0.5 (79.8,81.9)
GNB	41.0±0.9 (38.8,42.6)	90.1±0.2 (89.7,90.4)	45.3±6.6 (33.1,57.2)	41.0±0.9 (38.8,42.6)	33.5±1.3 (30.7,35.9)	73.6±0.5 (72.8,74.6)
KNN	61.2±1.5 (58.4,64.0)	93.5±0.2 (93.0,94.0)	60.7±1.6 (57.6,63.9)	61.2±1.5 (58.4,64.0)	60.3±1.5 (57.4, 63.2)	88.9±0.6 (72.8,74.6)
Best model	Random forest	Random forest	Random forest	Random forest	Random forest	Random forest

Note: Results are reflective of the model built on the data from the application of bootstrapping and is aggregated from the 1000-fold iterations of the 1000 bootstrapped re-sampling. Values within the brackets indicate the 95% CI lower and upper values.

Abbreviations: AUC-ROC, area under the receiver operating characteristic curve; CI, confidence interval; GNB, Gaussian Naïve Bayes; KNN, k-nearest neighbour; SVC, support vector classification.