



Machine learning prediction of susceptibility to visceral fat associated diseases

M. Aldraimli¹ · D. Soria² · J. Parkinson³ · E. L. Thomas³ · J. D. Bell³ · M. V. Dwek³ · T. J. Chausalet¹

Received: 16 January 2020 / Accepted: 2 June 2020 / Published online: 31 July 2020
© The Author(s) 2020

Abstract

Classifying subjects into risk categories is a common challenge in medical research. Machine Learning (ML) methods are widely used in the areas of risk prediction and classification. The primary objective of such algorithms is to use several features to predict dichotomous responses (e.g., healthy/at risk). Similar to statistical inference modelling, ML modelling is subject to the problem of class imbalance and is affected by the majority class, increasing the false-negative rate. In this study, we built and evaluated thirty-six ML models to classify approximately 4300 female and 4100 male participants from the UK Biobank into three categorical risk statuses based on discretised visceral adipose tissue (VAT) measurements from magnetic resonance imaging. We also examined the effect of sampling techniques on the models when dealing with class imbalance. The sampling techniques used had a significant impact on the classification and resulted in an improvement in risk status prediction by facilitating an increase in the information contained within each variable. Based on domain expert criteria the best three classification models for the female and male cohort visceral fat prediction were identified. The Area Under Receiver Operator Characteristic curve of the models tested (with external data) was 0.78 to 0.89 for females and 0.75 to 0.86 for males. These encouraging results will be used to guide further development of models to enable prediction of VAT value. This will be useful to identify individuals with excess VAT volume who are at risk of developing metabolic disease ensuring relevant lifestyle interventions can be appropriately targeted.

Keywords Machine learning · Imbalanced learning · UK biobank · Random under sampling · Synthetic minority over-sampling technique (SMOTE) · Visceral fat

1 Introduction

Real-world data are often imbalanced and lack uniform distribution across classes. Classification of imbalanced datasets is

a significant challenge across both industrial and research domains [1]. There are multiple approaches to tackle class imbalance [2], of which data enrichment is the most straightforward. Other more sophisticated methods include varied sampling techniques [3], cost-sensitive learning [4, 5], feature selection; more complex strategies include meta learning [6], combining classifiers [7], and algorithmic modifications [8].

When resampling methods are applied, questions over their suitability are often raised [9]. For example: is the new resampled dataset representative of the population in relation to the response variable? Is it acceptable to artificially generate synthetic data of class subjects when training Machine Learning (ML) classification models? It has been argued that by using sampling methods, the original class ratio is lost during the training process and that this affects the accuracy metrics [10]. Similarly, training ML models with synthetic data may compromise accuracy measures by deceiving the process of cross-validation sampling [11].

✉ M. Aldraimli
m.aldrainli@my.westminster.ac.uk

D. Soria
d.soria@kent.ac.uk

T. J. Chausalet
chausst@westminster.ac.uk

¹ School of Computer Science and Engineering, University of Westminster, London W1W 6UW, UK

² School of Computing, University of Kent (Medway), Chatham Maritime ME4 4AG, UK

³ School of Life Sciences, University of Westminster, London W1W 6UW, UK

In this paper, we compare the classification performance of six ML algorithms (Naïve Bayes, Logistic Regression, Artificial Neural Network, Decision Tree, Logistic Model Tree, and Random Forest) in predicting discretised visceral fat ranges associated with the development of long-term diseases in a multiclass classification problem. The new models were built using Random Under Sampling (RUS) [8] and Synthetic Minority Over Sampling Technique (SMOTE) [12] sampling techniques applied to highly imbalanced training data (in the female cohort case), and on less severe imbalance (in the male cohort case). This study suggests the most suitable models meeting the domain experts' success criteria. The data imbalance characteristic causing the transition in classifier training performance was monitored visually by Adaptive Projection Analysis (APA) [13] and numerically via Information Gain (IG) attribute evaluation [14, 15].

The deployment of machine learning modelling in this study aims at tackling a long-term real-world disease burden; Obesity affects an increasing number of adults in the UK [16], with obesity-associated changes in adipose tissue (AT) predisposing to metabolic dysregulation [17] and other disorders. Distribution of AT, in particular the accumulation of visceral adipose tissue (VAT) and liver fat, is a critical factor in determining susceptibility to diseases [18, 19]. Excess VAT and liver fat play a significant role in the pathogenesis of type 2 diabetes, dyslipidaemia, hypertension and cardiovascular disease [20].

Current strategies for the treatment of obesity and its associated co-morbidities have focused on lifestyle improvements [21, 22]. Such a focus aims to reduce VAT and liver fat, via calorie restriction and/or exercise, the impact of which are associated with improved insulin sensitivity, decreased blood pressure and lower circulating lipid levels [17, 23, 24]. Large scale analysis of the compartmental distribution of AT is often limited due to the expense and time required to employ the requisite imaging techniques. The UK Biobank (UKBB) provides a comprehensive means of assessing the relationship between body composition and lifestyle in a large population-based cohort of adults. Having such a large dataset could increase the presence of a pattern in the data, without it machine learning algorithm can't sufficiently learn to produce effective results.

The primary goal of this study is to identify the best models to predict VAT levels in a cohort of female and male individuals from the UKBB. The study is a cross sectional assessment of 4327 female and 4126 male individuals from the UKBB multimodal imaging cohort [25], aged 40–70 years and scanned chronologically between August 2014 and September 2016.

The paper is structured as follows: In Section 2, the methodology, methods and approaches used in this study are presented. In Section 3, the experimental design is shown. The

results are documented in section 4, with the discussion and conclusions Sections 5 and 6.

2 Methodology

For VAT prediction, multi-class ML classification models were applied to predict susceptibility to disease (risk) based on the discretised amount of VAT. Two groups of 2292 female and 2191 male subjects were used to train six ML algorithms using 10-fold cross-validation in three different scenarios. In relation to their cohort, the trained models were tested on two new groups of external data of 2035 and 1935 female and male cases, respectively. Figure 1 shows the methodology: multiple imbalanced datasets with the same predictor variables were modified with sampling techniques and used for modeling using the six ML algorithms. Selected performance metrics of the models were compared after training in the evaluation phase. IG was monitored for all predictor variables at every stage.

2.1 Data collection protocol

This cross-sectional study includes data from 8453 individuals included in the UKBB multimodal imaging cohort. The UKBB had approval from the North West Multi-Centre Research Ethics Committee (MREC), and written consent was obtained from all participants before their involvement. The data was acquired through the UK Biobank Access Application number 23889. The age range for inclusion was 40–70 years, with exclusion criteria were: metal or electric implants, medical conditions that prohibited MRI scanning or planned surgery within 6 weeks before the scanning date. The subjects were scanned chronologically between August 2014 and September 2016. The visceral adipose tissue (VAT) volumes were acquired as part of the UKBB dataset.

Anthropometry measurements were collected at UKBB assessment centers; height was measured using the Seca 202 height measure (Seca, Hamburg, Germany). The average of two blood pressure measurements, taken moments apart, was obtained using an automated device (Omron, UK). Images were acquired at the UK biobank imaging Centre at Cheadle (UK) using a Siemens 1.5 T Magnetom Aera. The participants' height and weight were recorded before imaging screening which later was utilised to calculate the Body Mass Index (BMI).

For physical activity assessment data, a touchscreen questionnaire was used to collect information on sociodemographic characteristics and lifestyle exposures (<http://www.ukbiobank.ac.uk/resources/>). Specific questions on the frequency and duration of walking (UK biobank field ID: 864, 874), moderate physical activity (884, 894) and vigorous physical activity (904, 914) events allowed the

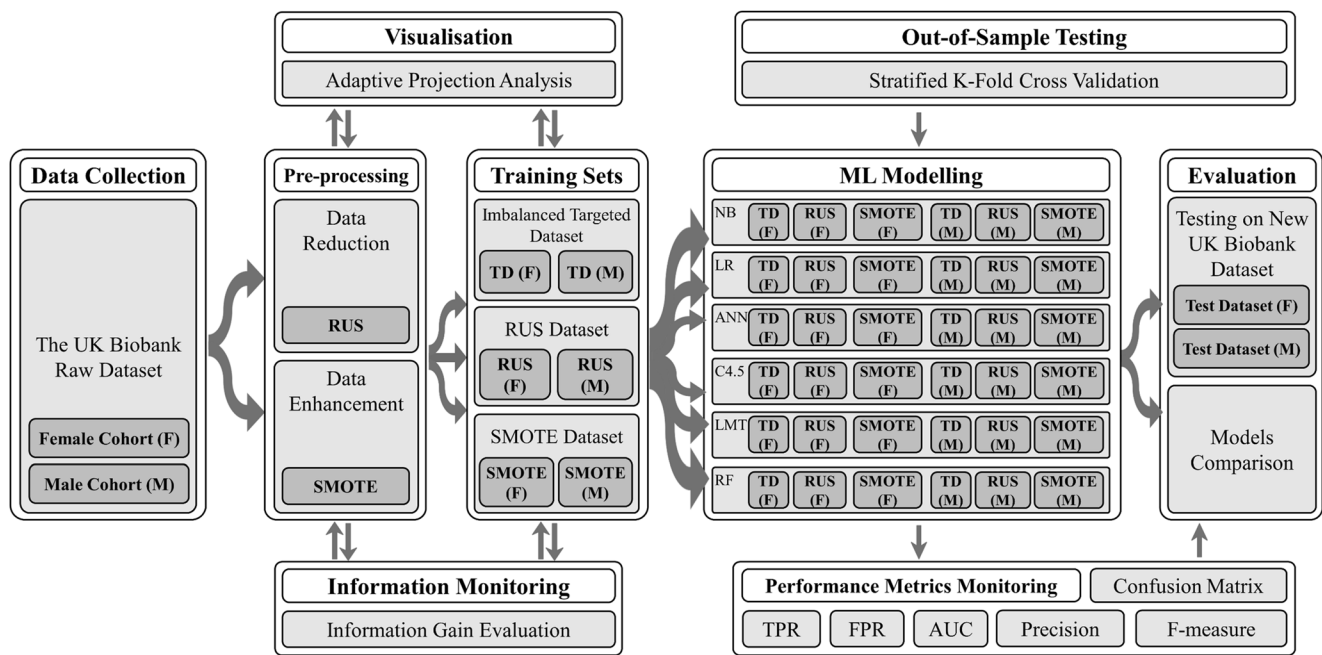


Fig. 1 The methodology adopted in this work, showing the different steps followed. Where TD = Targeted dataset, RUS = Random Under Sampling, SMOTE = Synthetic Minority Oversampling Technique, ML = Machine Learning, NB = Naïve Bayes, LR = Logistic Regression,

ANN = Artificial Neural Network, C4.5, LMT = Logistic Model Tree, RF = Random Forest, TPR = true-positive rate, FPR = false-positive rate, AUC = Area under receiver operator characteristic curve

calculations of metabolic equivalent-minutes per week (MET-min/week) for each individual. Participants were excluded from the calculations and analysis if they selected ‘prefer not to answer’ or ‘do not know’ to any of the possible six questions on physical activity used to calculate the MET score ($n = 868$).

2.2 Information gain evaluation algorithm

Information and entropy levels within independent variables were monitored using an Information Gain Attribute Evaluator Algorithm [15]. This algorithm evaluates the worth of each attribute by measuring information gained with respect to the class in combination with a ranker algorithm that ranks the attributes by their influence on the class [14, 15, 26].

2.3 Adaptive projection analysis (APA)

APA uses a linear projection to display high dimensional data into 3-dimensions by allowing the user to drag points in an interactive scatter plot to find new views [13]. These views indicate the classes which can be separated, the attribute combinations which are most associated with each class, the outliers, the sources of error in the classification algorithms, and the existence of clusters in the data [27].

2.4 Data preprocessing

Pre-processing (preparation) steps are applied to the training dataset depending on various observed characteristics within the data, i.e., dataset dimensions, units of measurements and distribution. Preprocessing the data aims to change classifiers behavior in the modeling phase. Some forms of classifier behavior changes are adding bias towards a response group, adding more weight to a feature and taking a classification cost into account. Dataset class imbalance requires the training dataset to undergo resampling processes. Resampling methods are one of many different approaches known to improve imbalanced learning [2–8]. The application of resampling techniques enhances the training dataset in the form of data reduction or enrichment. The following two resampling techniques were applied in this study.

Random Undersampling (RUS) This approach consisted of selecting a subset of the majority class to balance the data [8]. In this approach (Fig. 2), some of the majority class records were removed at random. However, it was recognised that deleting records could lead to loss of important information or patterns which may have been relevant to the learning process [28]. Denoting the majority class L and the minority class S , r was defined as the ratio between the size of the minority and majority classes [3]. We performed random under-sampling of L to achieve a balanced ratio of $r = 1$.

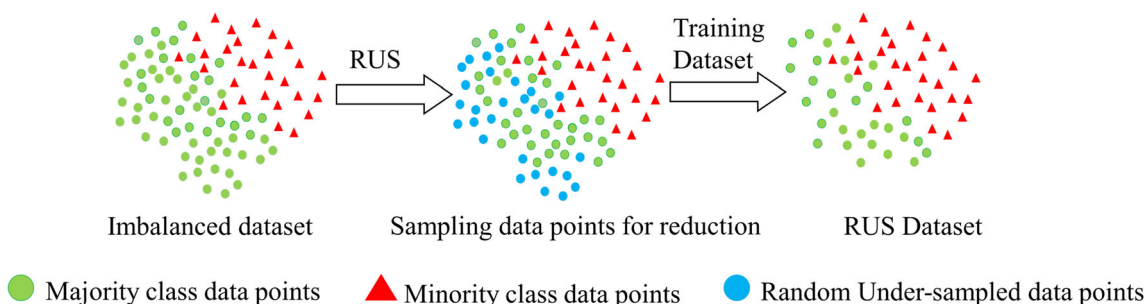


Fig. 2 Illustration of random undersampling technique

The imbalanced r ratios before RUS were r (females) = 0.14 and r (males) = 0.43.

Synthetic minority oversampling technique (SMOTE) SMOTE is an over-sampling technique developed by Chawla [12]. It aims to enhance the minority class by creating artificial examples in the minority class (Fig. 3). For each data point x in S (the minority class), one of its k -nearest neighbours ($k = 5$) was identified. The k neighbours were randomly selected, and artificial observations were generated and spread in the area between x and the nearest neighbours. These synthetic points were added to the dataset in class S . The artificial generation of the data points differed from the multiplication method [16] to avoid the problem of overfitting.

2.5 ML Modelling

The classification process in this study uses a predictive learning function that classifies an observation into one of three predefined (labeled) classes. The six ML classification algorithms selected for this study have different learning schemes such as graphical model-based classifiers, curve-fitting algorithms, tree-based techniques and ensemble learners. The use of such a variety is to examine the effect of different learning schemes on the final results.

Naïve Bayes (NB) A probabilistic graphical model-based machine learning classifier used for classification tasks. The foundation of the classifier is the Bayes Theorem [29]. It also assumes that predictor variables are independent and that all

predictor variables have an equal effect on the response outcome. Despite the simplified assumptions of Naïve Bayes classifiers, they were reported to be useful in complex real-world situations [30].

Logistic Regression (LR) LR is a deterministic curve-fitting technique which produces a probability-based model that accounts for the likelihood of an event occurring (the value of the class variable) depending on the values of the predictors (categorical or numerical) [31, 32].

Artificial neural network (ANN) ANNs are used to fit observed data, unusually high dimensional datasets characterised by noise and missingness (pollution). Neural networks comprise elementary autonomous computational units, known as neurons. Neurons are interconnected via weighted connections and organised in layers (an input layer, hidden layers and an output layer). In this study, a Multi-Layer Perceptron (MLP) ANN with a sigmoid activation function was used, [17] as a curve-fitting classifier.

Decision tree (C4.5) The C4.5 algorithm is used in data mining as a Decision Tree Classifier which generates a decision, based on a sample of data. In this method, a new data point is predicted (classified) via a series of tests to determine its class. The tests hierarchically assemble a tree of decisions, hence ‘decision tree’ [15, 33, 34].

Logistic model tree (LMT) LMT is an ensemble model with a tree structure but with LR functions at the leaves level. The

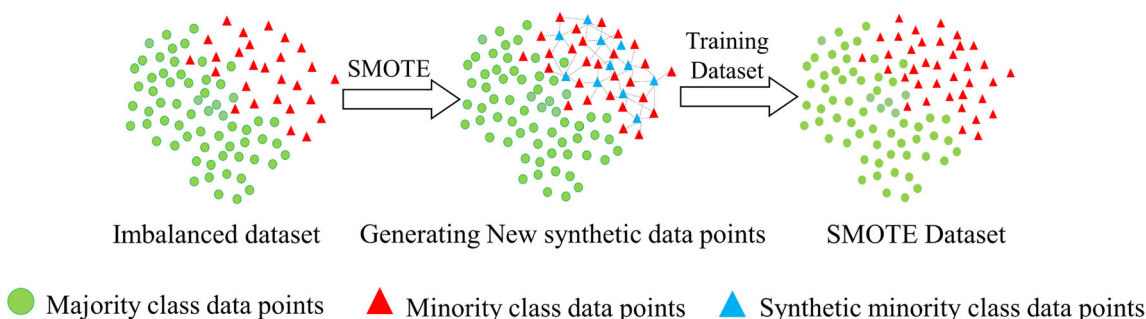


Fig. 3 Illustration of synthetic minority oversampling technique

LMT structure comprises a set of non-terminal nodes and a set of leaves (terminal nodes). LMT is designed to adapt to small data subsets where a simple linear model offers the best bias-variance trade-off [31].

Random Forest (RF) RF is another ensemble learner and a generalisation of standard decision trees proposed by Breiman based on bagging (Bootstrap Aggregation) from a single training set or random not pruned decision trees [18]. Bootstrap Aggregation is used to combine the predictions of the individual trees [19].

All the six methods used for this study were implemented in Weka [35] (with default parameters settings), with the C4.5 using the J48 implementation.

2.6 Out-of-sample testing

Out of sample testing is also known as cross-validation [11] aims to test the model’s capability of predicting (classifying) new data that was not used for training it. Cross-validation provides an insight on how the model will generalize to a new unknown dataset.

Cross-validation can be performed in several rounds (folds) *k* (see Fig. 4). A fold of cross-validation involves partitioning a sample of data into subsets, performing the model’s training on one subset, and testing the model on the other subset. Where multiple rounds of cross-validation are performed using different partitions, the test results are averaged over the folds to estimate the model’s classification performance.

2.7 Model evaluation

The following measures were chosen to evaluate the performance of each model: accuracy (later reported as Correctly Classified Instances ratio or ‘CCI’) true-positive rate (TPR, also known as sensitivity or recall), specificity, false-positive rate (‘FPR’), precision (‘Prcn’), area under the receiver operator characteristic curve (‘ROC’), and F-measure (‘F-m’) [36–38]. The latter is a harmonic mean of precision and recall. Practically, a high F-measure value indicates that both recall and precision are high, meaning fewer subjects misdiagnosed with a disease or risk of disease. The F-measure is essential to assess the model performance when classifying very imbalanced data [37].

True positive (TP), true negative (TP), false positive (FP) and false negative (FN) TP is the number of correctly classified instances in a risk group (class), TN is the number of correctly classified instances in other groups, FP also known as false alarm or type-I error is the number of incorrectly classified instances in healthy and moderate groups as at risk and FN also known as type-II error is the number of incorrectly classified instances in a risk group.

Accuracy (CCI) Model accuracy is the ratio of all examples in a dataset which were correctly classified. Also known as Correctly Classified Instances ratio CCI.

$$CCI = \frac{TP + TN}{TP + TN + FP + FN}$$

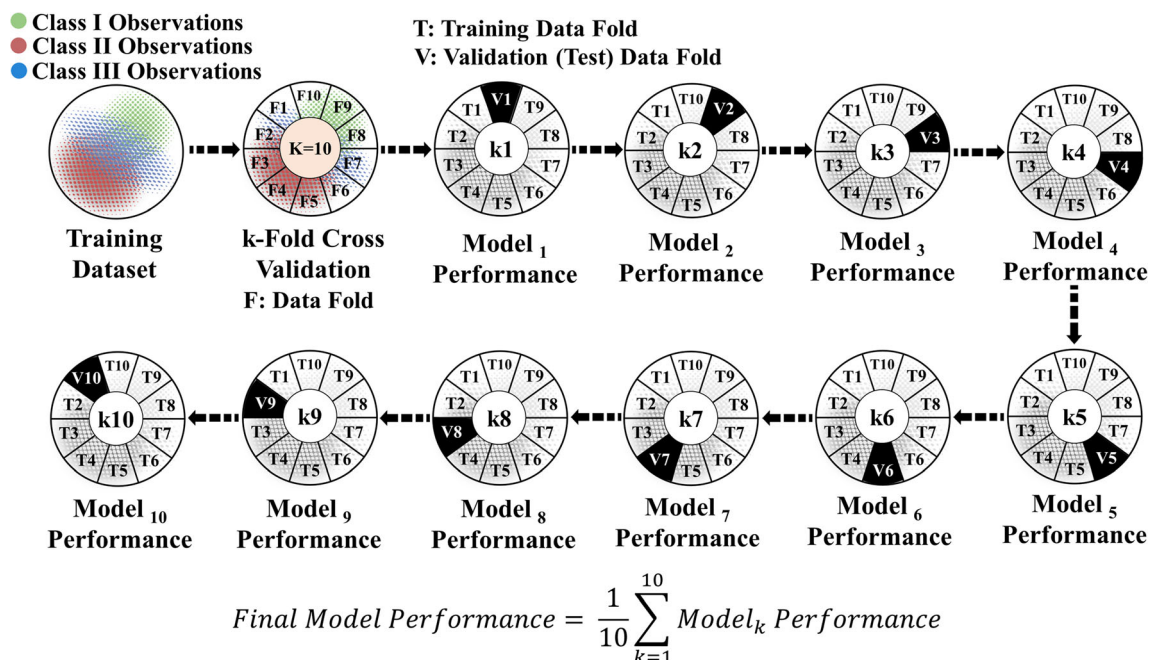


Fig. 4 Illustration of k-fold cross-validation

Recall Also, known as sensitivity or true positive rate (TPR); assume having a group whose members are at risk of a disease, the true positive rate (TPR) in a risk class (group) is the ratio of number of subjects who were predicted correctly as at risk to the total number of subjects of risk group (both predicted correctly and incorrectly).

$$\text{Recall} = \frac{TP}{TP + FN}$$

Specificity Specificity is also known as true negative rate (TNR); Assume having a group who are risk-free of a disease (Healthy class), the true negative rate in a risk-free class is the ratio of number of subjects were predicted correctly as at risk-free to the total number of subjects of risk-free group (both predicted correctly and incorrectly).

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Precision (Prcn) This performance metric is also known as positive predictive value (PPV) which is the ratio of true positive (TP) predictions to all correctly and incorrectly predicted positive predictions (TP + FP)

$$\text{Precision} = \frac{TP}{TP + FP}$$

F-measure (F-m) Also known as the harmonic mean of precision and recall. Practically, a high F-measure value indicates that both recall and precision are high which means the less subjects are misdiagnosed with a disease or risk. This metric is important to assess the model performance when classifying minority class.

$$F\text{-measure} = 2 * \frac{\text{Prcn} * \text{Recall}}{\text{Prcn} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

Area under the curve (AUC) The area under the receiver operator characteristic (ROC) curve is a method to comparing classifiers performances. From the ROC graph example in Fig. 5, it is possible to obtain an overall evaluation of quality. AUC is the fraction of the total area which falls under the ROC curve. FPR is the false positive rate. The AUC is calculated by

$$AUC = \int_{x=0}^1 TPR(FPR^{-1}(x)) dx$$

The AUC value is in the range of 0.5 to 1, where 0.5 denotes a bad performing classifier and 1 denotes an excellent

performing classifier. In medical diagnosis, experts seek very high AUC value.

Confusion matrix It is also known as the error matrix. Figure 6 shows all outcomes of the classification formulated into $m \times m$ matrix. The confusion matrix layout is useful when visualising the performance of a classification algorithm. Each row of the matrix represents the predicted instances in a class while each column represents the actual instances in a class.

3 Experimental design

VAT- related disease susceptibility was based on the following MRI response labels (Fig. 7): Healthy, Moderate and Risk defined according to VAT volume. In females; VAT volume of ≤ 2 was deemed ‘Healthy’ (H); VAT volume > 2 but ≤ 5 was classed as ‘Moderate’ (M); VAT volume > 5 was classified as ‘Risk’ (R) [39]. In males; VAT volume ≤ 3 was deemed ‘Healthy’ (H); VAT volume was > 3 but ≤ 6 , was classed as ‘Moderate’ (M); VAT volume > 6 , was classed as ‘Risk’ (R) [39]. The training datasets contained ten data variables reported in Table 1, with the VAT in liters being the class determination response variable. All nine predictor variables in Table 1 were selected as input features by domain experts based on their correlation with VAT prediction in previous studies which are discussed in section 5.

The UK Biobank Physical Activity Index (UKBB PAI or PAI) was created by domain experts [40] using data collected during physical activity assessment; comprising a total of 27 outcomes, 23 outcomes reflecting activity and four reflecting inactivity (see Table 2). An individual’s response to questions was scored with values between -1 and $+1$ and combined cumulatively to give a final score. With an increasingly negative score implying a progressively unhealthier phenotype. For binary variables 0 indicated absence of the parameter, 1 the presence.

Targeted dataset (TD) The TD was the first dataset modelled. The TD contained 2292 female and 2191 male records, from the UKBB cohort. Table 1 shows the summary statistics of all TD’s variables. The TD was highly imbalanced in the female cohort and less severely imbalanced in the male cohort in relation to records numbers per class: In the females’ TD class H had 1002 subjects, class M had 1128 subjects, and class R contained only 162 subjects. In the males’ TD class H had 489 subjects, class M had 1125 subjects, and class R contained 577 subjects. The class imbalance of TD can be observed via APA visualisation in Fig. 8.

Random under-sampled (RUS) dataset This dataset was a reduced subset of TD. A subset of each majority class was randomly removed to balance the data. As a result of applying

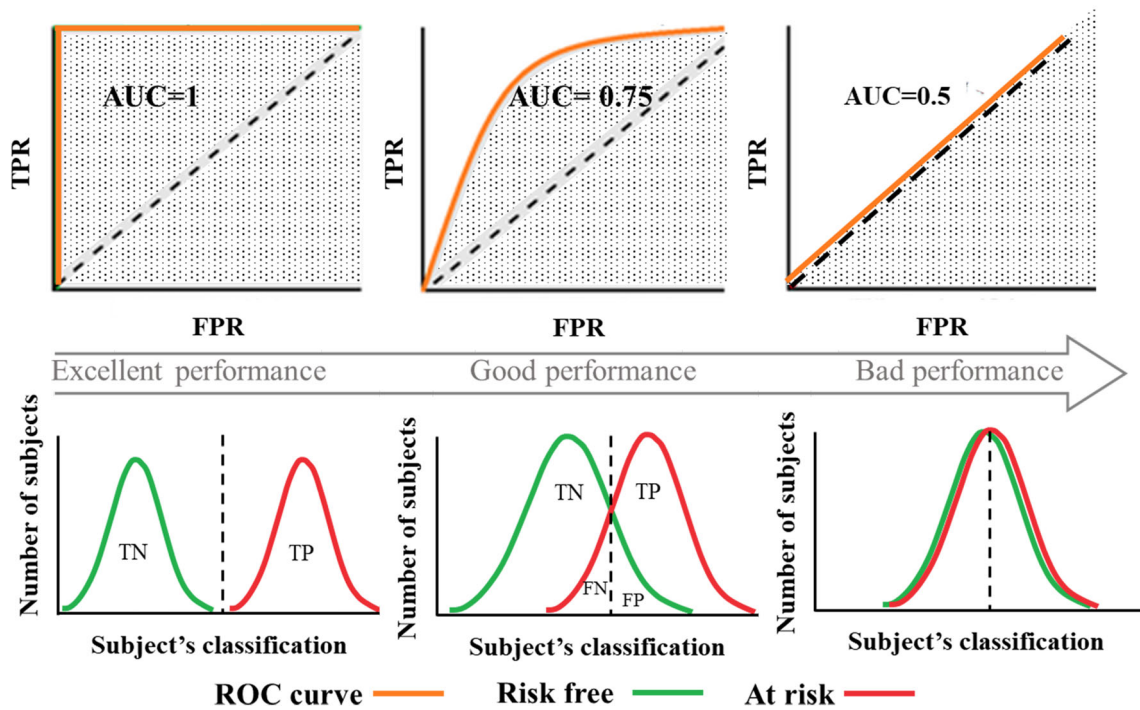


Fig. 5 Illustration of AUC classification metric

RUS to the females’ TD, each of the H, M and R classes ended up with 162 subjects. While in the males’ TD each of the H, M and R classes ended up with 489 subjects. The effect of RUS can be observed APA visualisation in Fig. 8.

Synthetic minority over-sampled (SMOTE) dataset This dataset was obtained as a result of applying SMOTE to the numeric data variables of TD. By doing so, the three VAT classes became more closely balanced. In the female cohort, class H had 1002 subjects, class M had 1128 subjects, and class R contained 1296 subjects. In the male cohort, class H had 1125 subjects, class M had 1125 subjects and class R contained 1125 subjects. The effect of SMOTE can be observed via APA visualisation in Fig. 8.

Fig. 6 Classification confusion matrix layout

		<i>Predicted Class</i>		
		<i>Risk Class</i>	<i>Risk-free Class</i>	
<i>Actual Classification</i>	<i>Risk Class</i>	True Positive (TP)	False Negative (FN)	Recall $\frac{TP}{TP + FN}$
	<i>Risk-free Class</i>	False Positive (FP)	True Negative (TN)	Specificity $\frac{TN}{TN + FP}$
		Precision $\frac{TP}{TP + FP}$	Negative Predictive Value $\frac{TN}{TN + FN}$	

IG Evaluation Algorithm was used to measure the information levels for independent variables in relation to the class variable. The measurement and ranking of IG in each independent variable in TD, RUS and SMOTE training sets are presented in Section 3.

The Test Dataset The ML models were tested on a new group of 2035 females from the UKBB female cohort and a new group of 1935 males from the UKBB male cohort. The same ten variables as per the training datasets, were used to test all models. Table 3 shows their summary statistics. Like the TD, the female Test Dataset was also highly imbalanced: class H had 823 subjects, class M had 1039, and class R contained

Fig. 7 The response labels of VAT-related disease susceptibility

a) Females VAT MRI response labels		b) Males VAT MRI response labels	
VAT volume > 5.0 Litres	Risk	VAT volume > 6.0 Litres	Risk
2.0 < VAT volume ≤ 5.0 Litres	Moderate	3.0 < VAT volume ≤ 6.0 Litres	Moderate
VAT volume ≤ 2.0 Litres	Healthy	VAT volume ≤ 3.0 Litres	Healthy

only 173 subjects. The males Test Dataset was less severely imbalanced: class H had 468 subjects, class M had 906, and class R contained 561 subjects.

4 Results

4.1 ML models training results

From the confusion matrices in Table 4, the model training accuracies for the female cohort, presented as Correctly Classified Instances ratio (CCI) or True Positive Rate (TPR)

of all methods were computed, they showed that resampling methods resulted in an improvement in CCI compared to the original TD. When the performance of the LR, ANN, C4.5 and RF models for the female cohort was evaluated, it was apparent that the RUS dataset was poorer than when the TD data set was used, Fig. 9.

The AUC for each of the trained models were in the range of 0.783 (for RF on SMOTE) to 0.96 (for C4.5 on TD). These values indicate that the trained models did not sacrifice much precision to achieve a good recall value on the observed data points. The RF model achieved the highest TPR (0.850) when trained on the SMOTE dataset, while the C4.5 model achieved the lowest TPR (0.714) when trained on the RUS dataset.

Table 1 Descriptive statistics of variables in the Targeted Dataset (TD)

Female Cohort ($n = 2292$)			
Numeric selected dataset variables	Median	Mean	(Min, Max)
Response variable			
Visceral adipose tissue volume (VAT in litres)	2.2	2.5	(0.1, 9.7)
Predictor variables			
Waist Circumference (WC in cm)	80.0	81.6	(55.0, 126.0)
Pre-imaging Weight (W in Kg)	66.0	68.3	(42.0, 128.0)
BMI (in kg/m ²)	24.8	25.7	(15.5, 48.0)
Hip circumference (HC in cm)	100.0	100.9	(77.0, 147.0)
Standing height (H in cm)	163.0	163.0	(141.0, 194.0)
Systolic blood pressure (SBP in mmHG)	133.0	134.5	(87.0, 225.0)
Diastolic blood pressure (DBP in mmHG)	77.0	77.8	(45.0, 120.0)
Physical Activity Index (PAI)	0.5	0.6	(−12.0, 15.5)
Age at recruitment (AGE in years)	55.0	54.6	(40.0, 70.0)
Male Cohort ($n = 2191$)			
Response variable			
Visceral adipose tissue volume (VAT in litres)	5.6	4.7	(0.35, 9.63)
Predictor variables			
Waist Circumference (WC in cm)	102.0	92.5	(66.0, 138.0)
Pre-imaging Weight (W in Kg)	104.0	82.5	(53.0, 155.0)
BMI (in kg/m ²)	33.0	26.6	(18.0, 48.0)
Hip circumference (HC in cm)	116.5	101.1	(83.0, 150.0)
Standing height (H in cm)	176.0	176.1	(152.0, 200.0)
Systolic blood pressure (SBP in mmHG)	159.0	141.9	(99.0, 219.0)
Diastolic blood pressure (DBP in mmHG)	84.5	80.8	(51.0, 118.0)
Physical Activity Index (PAI)	3.0	0.5	(−12.0, 18.0)
Age at recruitment (AGE in years)	55.0	56.4	(40.0, 70.0)

Table 2 UK Biobank outcomes used in creating the physical activity index

UK BB ID	Outcome	Units
816	Job involves heavy lifting	Categorical
864	Days/week walked 10+ minutes	Days/Week
874	Duration of Walks	Minute/Day
884	Days/week moderate physical activity 10+ minutes	Days/Week
894	Duration of moderate activity min	Min/Day
904	Days/week vigorous physical activity 10+ minutes	Days/Week
914	Duration of vigorous activity	Minute/Day
924	Usual walking pace	Categorical
943	Frequency of stair climbing in last 4 weeks	Categorical
971	Frequency of walking for pleasure in last 4 weeks	Categorical
981	Duration of walking for pleasure	Categorical
991	Frequency of strenuous sports in last 4 weeks	Categorical
1001	Duration of strenuous sports	Categorical
1011	Frequency of light DIY in last 4 weeks	Categorical
1021	Duration of light DIY	Categorical
2624	Frequency of heavy DIY in last 4 weeks	Categorical
2634	Duration of heavy DIY	Categorical
3637	Frequency of other exercises in last 4 weeks	Categorical
3647	Duration of other exercises	Categorical
6164	Types of physical activity in past 4 weeks	Categorical
104,900	Time spent doing vigorous physical activity	Categorical
104,910	Time spent doing moderate physical activity	Categorical
104,920	Time spent doing light physical activity	Hours
806	Job involves mainly standing or walking	Categorical
1070	Time spent watching television	Hour/Day
1080	Time spent using computer	Hour/Day
1090	Time spent driving	Hour/Day

By observing the confusion matrices for all models after training on all the TD and RUS datasets, it is clear that the number of incorrectly classified instances for class R highly decreased for the models trained on the RUS dataset compared to those trained on the TD. However, when evaluating the minority class accuracy performance in Fig. 10, it is notable that all trained models benefitted from the sampling methods, exhibiting consistent TPR improvement for class R in each model.

The accuracies (CCI) of the models for the male cohort were calculated from Table 4. SMOTE resampling resulted in a consistent improvement in CCI as compared to the original TD. SMOTE resampling resulted in a consistent improvement in CCI as compared to the original TD. The training performance of all models for the male cohort using the RUS dataset was reduced compared to the same algorithms trained on the TD. The AUC for each of the trained models were in the range of 0.729 (for C4.5 on TD) to 0.923 (for RF on SMOTE). These values indicate that the trained models did not sacrifice a lot of precision to obtain a good recall value on

the observed data points. The RF model trained on the SMOTE dataset achieved the highest TPR (0.793), while the C4.5 model trained on the RUS dataset achieved the lowest TPR (0.631).

Examination of the confusion matrices for all models trained on the TD vs the RUS datasets demonstrated that the number of subjects incorrectly classified as class H instead of class R increased for models trained on the RUS dataset compared with those trained on the original TD despite the removal of 88 subjects from the original R group as a result of RUS. The number of correctly classified instances for class H increased. However, when evaluating class R accuracy performance (see Fig. 10), it is notable that all trained models benefitted from the sampling methods, exhibiting consistent TPR improvement for class R in each model.

4.2 Models test results

The models derived above were tested on a further dataset (female, $n = 2035$; male $n = 1935$). When the CCI values for

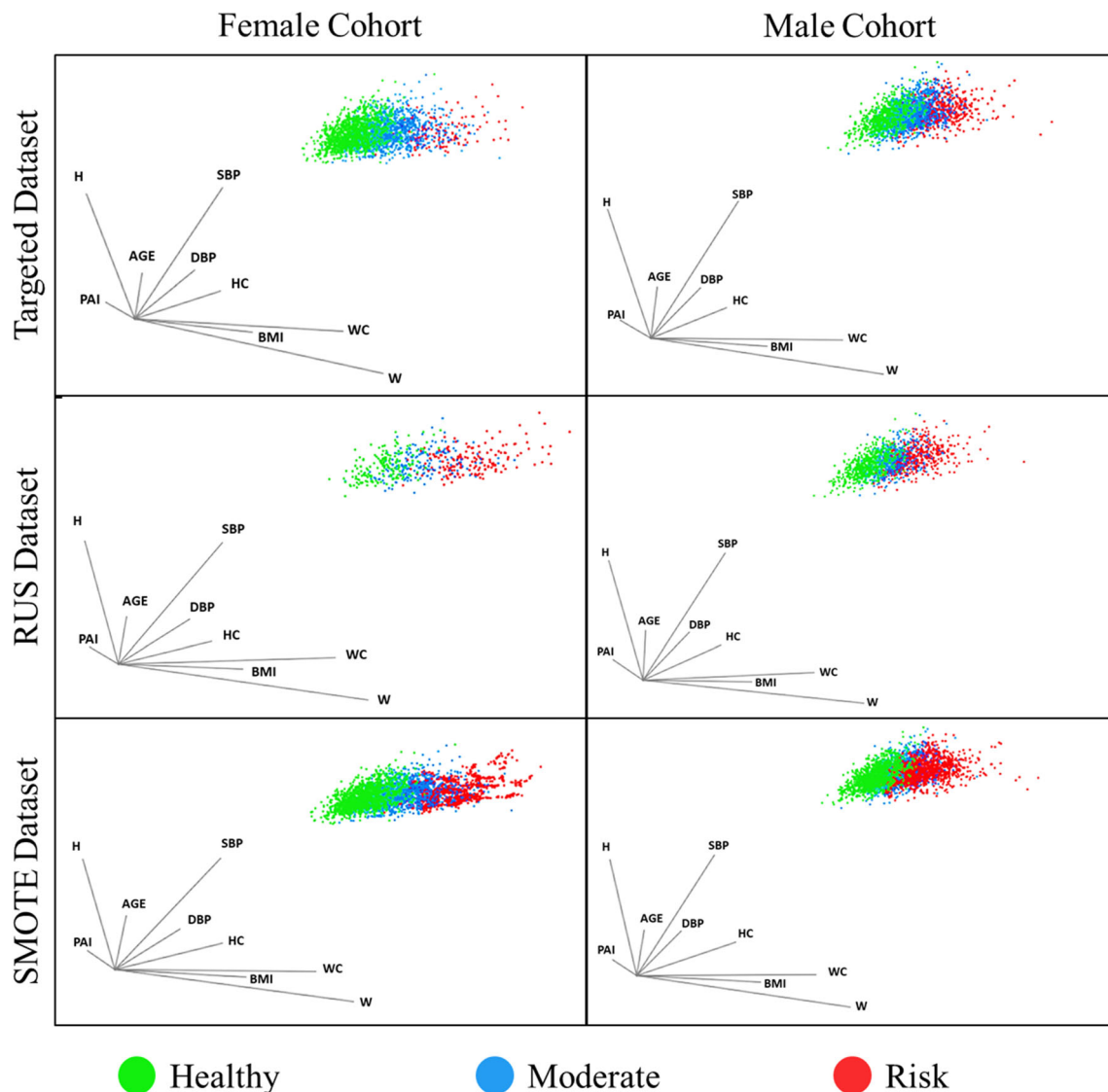


Fig. 8 Adaptive projection visualisation of Targeted Dataset, Random Under Sampled dataset and SMOTE dataset variables

all models were compared using the female cohort, the CCI decreased to a maximum degradation of 6.2% when testing the C4.5 model trained on the RUS dataset against the same algorithm trained on the original TD. LMT model built with SMOTE dataset achieved an overall test accuracy improvement of 6.83% when compared to TD.

In the female cohort (see Fig. 11) RF models achieved the best TPR of 0.770 when trained on the TD dataset. LMT model achieved the least TPR of 0.681 when trained on the TD dataset. The ROC area across all tested models ranged between 0.786 (for C4.5 on SMOTE dataset) and 0.889 (for LR on TD). These values indicate hardly any loss of precision whilst achieving a good recall value on the observed data points. For evaluating risk class, R, the TPR performance (Fig. 10) classified the risk group with the highest level of 0.798 was achieved by RF on RUS. RF also achieved the greatest TPR improvement in test with a difference of 0.463

between RUS and TD. NB ranked last, with just 0.121 in minority class TPR improvement between NB on SMOTE and TD. These results can be visualised in the confusion matrixes in Table 4. The RF model trained on SMOTE correctly classified the highest number of instances (138 of the original 173) in class R. The model which performed the worst in TPR performance for the class R was C4.5 trained on TD, which only correctly classified 43 instances.

In male cohort subjects; when comparing the CCI for all models, CCI decreased with a maximum degradation of 11.9% when testing the RF model trained on the SMOTE dataset compared to the same model built on the TD. All models built on the TD showed an overall model accuracy improvement on test datasets, the highest model accuracy improvement 4.0% was achieved with C4.5 model trained on TD dataset when compared to all other models. The models' overall accuracy improvements in test were also observed for NB,

Table 3 Descriptive statistics of variables in the test set

Female Cohort (<i>n</i> = 2035)			
Numeric test dataset variables	Median	Mean	(Min, Max)
Response variable			
Visceral adipose tissue volume (VAT in litres)	2.4	2.7	(0.2, 10.0)
Predictors variables			
Waist Circumference (WC in cm)	80.0	81.6	(55.0, 142.0)
Pre-imaging Weight (W in Kg)	67.0	68.7	(39.0, 136.0)
BMI (in kg/m ²)	25.2	25.9	(14.4, 54.5)
Hip circumference (HC in cm)	100.0	101.3	(73.0, 156.0)
Standing height (H in cm)	163.0	162.7	(145.0, 195.0)
Systolic blood pressure (SBP in mmHG)	129.0	130.4	(87.0, 196.0)
Diastolic blood pressure (DBP in mmHG)	76.0	76.6	(45.0, 115.0)
Physical Activity Index (PAI)	0.0	0.1	(−12.5, 18.0)
Age at recruitment (AGE in years)	55.0	54.6	(40.0, 70.0)
Male Cohort (<i>n</i> = 1935)			
Numeric test dataset variables	Median	Mean	(Min, Max)
Response variable			
Visceral adipose tissue volume (VAT in litres)	7.2	4.9	(0.3, 14.1)
Predictors variables			
Waist Circumference (WC in cm)	101.0	93.3	(63.0, 139.0)
Pre-imaging Weight (W in Kg)	100.0	83.4	(50.0, 150.0)
BMI (in kg/m ²)	32.5	26.9	(17.0, 48.0)
Hip circumference (HC in cm)	109.5	101.4	(78.0, 141.0)
Standing height (H in cm)	178.5	175.8	(156.0, 201.0)
Systolic blood pressure (SBP in mmHG)	142.0	137.1	(75.0, 209.0)
Diastolic blood pressure (DBP in mmHG)	83.5	79.9	(47.0, 120.0)
Physical Activity Index (PAI)	2.8	0.5	(−12.0, 17.5)
Age at recruitment (AGE in years)	55.0	56.0	(40.0, 70.0)

LR and MLP models trained on RUS dataset with the greatest improvement of 1.2% on NB when compared to all models built with RUS dataset. All models built with SMOTE dataset suffered an overall model accuracy degradation in test except for NB overall accuracy that remained unchanged.

In the male cohort, it was observed that in test, LR models achieved the best TPR of 0.733 when trained on TD dataset (see Fig. 11). LMT model achieved the least TPR of 0.730 when trained on TD dataset. The ROC area across all tested models ranged between 0.753 (for C4.5 on SMOTE) and 0.864 (for LR on both TD and SMOTE, and LMT on TD). These values indicate that also, the tested models do not sacrifice much precision to obtain a good recall value on the observed data points.

When observing class R, the TPR performance results in Fig. 10 show that consistent improvements were made in classifying the risk group with the highest level of 0.836 achieved by LMT on RUS.

LMT also achieved the greatest TPR improvement in test with a difference of 0.164 between LMT on RUS and LMT on TD, while MLP ranked last, with just 0.05 in class R TPR improvement between MLP on RUS and TD. This comparison is demonstrated in the confusion matrixes in Table 4. The LMT model trained on RUS correctly classified the highest number of instances (469 of the original 561) in class R. The model which performed the worst in TPR performance for class R was NB trained on TD, which only correctly classified 361 instances.

The effect of using a variety of ML algorithms with different learning schemes is examined. At a model level, Fig. 8 shows a small difference between the minimum and the maximum TPR test performances per dataset in each cohort. In the females, tested TD, RUS and SMOTE models showed only differences of 0.1, 0.06 and 0.05 respectively between the highest and the lowest performing algorithms. A similar pattern is found in the

Table 4. All cohorts VAT Prediction Models Confusion Matrices

a Female Cohort VAT Prediction Confusion Matrices

	TD			RUS Dataset						SMOTE Dataset														
	Training			Test			Training			Test			Training			Test								
	H	M	R	H	M	R	H	M	R	H	M	R	H	M	R	H	M	R						
NB	855	147	0	H	720	103	0	H	143	19	0	H	742	80	1	H	856	146	0	H	721	102	0	H
	287	728	118	M	283	662	94	M	29	113	20	M	342	551	146	M	289	668	171	M	283	600	156	M
	0	50	112	R	2	69	102	R	0	14	148	R	2	54	117	R	0	231	1065	R	2	48	123	R
LR	888	169	0	H	698	125	0	H	186	26	0	H	707	115	1	H	884	167	1	H	699	123	1	H
	180	915	33	M	188	828	23	M	30	106	26	M	211	681	147	M	184	769	175	M	189	704	146	M
	0	89	73	R	1	104	68	R	0	25	137	R	1	44	128	R	0	127	1169	R	1	40	132	R
ANN	809	193	0	H	664	159	0	H	123	38	1	H	671	149	3	H	813	187	2	H	685	133	5	H
	177	907	44	M	152	875	12	M	32	93	37	M	234	678	127	M	162	783	183	M	186	701	152	M
	0	92	70	R	1	128	44	R	1	26	153	R	2	55	116	R	0	83	1213	R	2	50	121	R
C4.5	741	261	0	H	606	216	1	H	125	35	2	H	723	93	7	H	739	239	4	H	680	142	1	H
	151	922	55	M	140	879	20	M	44	90	28	M	299	561	179	M	170	819	139	M	224	710	105	M
	0	82	80	R	1	129	43	R	1	26	132	R	3	52	118	R	2	90	1204	R	2	83	88	R
LMT	765	246	1	H	637	185	1	H	134	28	0	H	727	95	1	H	818	184	0	H	676	145	2	H
	235	808	85	M	284	695	60	M	29	107	26	M	249	634	136	M	180	792	156	M	180	737	122	M
	0	91	71	R	3	117	53	R	0	27	135	R	1	39	133	R	1	75	1220	R	3	59	111	R
RF	823	179	0	H	679	144	0	H	130	32	0	H	694	127	2	H	811	191	0	H	681	142	0	H
	175	916	37	M	183	829	27	M	34	90	38	M	204	649	186	M	169	848	111	M	168	783	86	M
	0	90	72	R	1	114	58	R	0	24	138	R	3	32	138	R	0	43	1253	R	2	78	93	R

b Male Cohort VAT Prediction Confusion Matrices

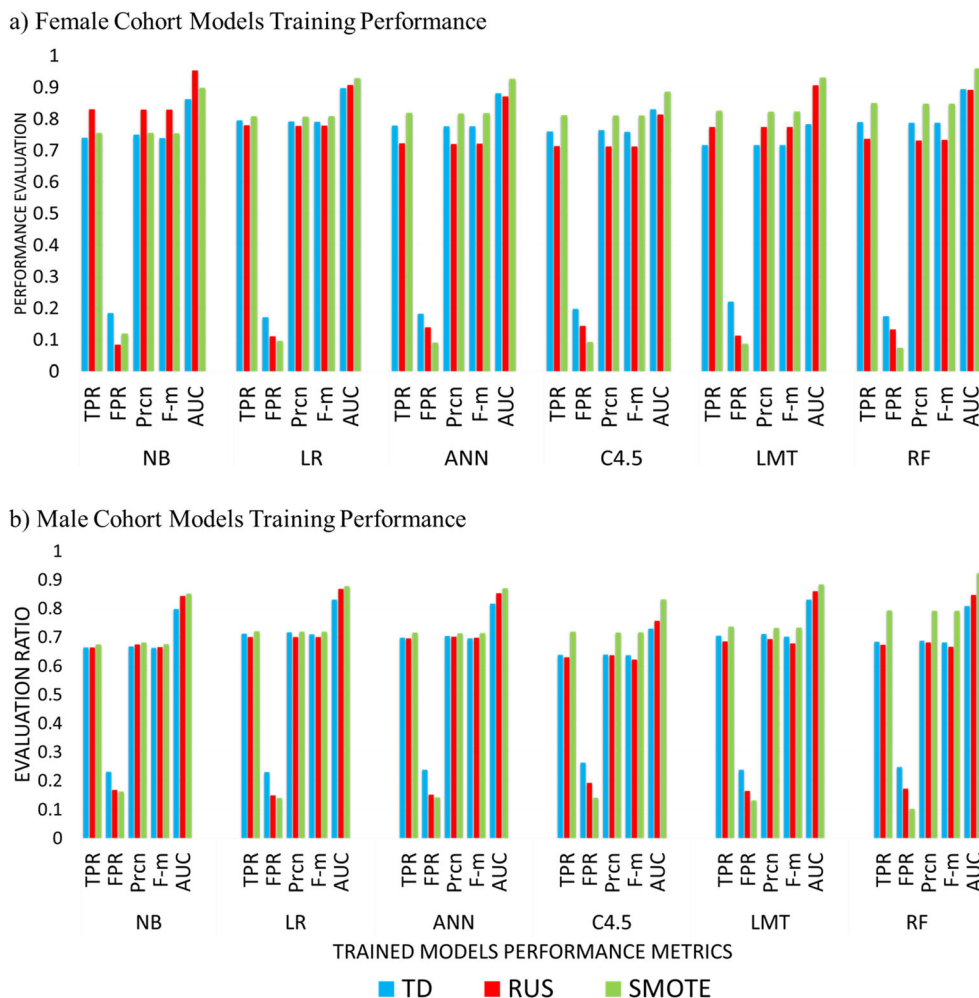
	TD			RUS Dataset						SMOTE Dataset														
	Training			Test			Training			Test			Training			Test								
	R	M	H	R	M	H	R	M	H	R	M	H	R	M	H	R	M	H						
NB	331	242	4	R	361	198	2	R	308	171	10	R	403	154	4	R	738	371	16	R	398	156	7	R
	143	762	220	M	117	637	152	M	86	278	125	M	162	528	216	M	200	620	305	M	160	526	220	M
	2	127	360	H	1	127	340	H	5	98	386	H	3	89	376	H	6	203	916	H	2	85	381	H
LR	350	226	1	R	383	177	1	R	372	114	3	R	451	108	2	R	862	258	5	R	451	108	2	R
	111	901	113	M	108	739	59	M	113	272	104	M	193	539	174	M	237	650	238	M	193	544	169	M
	0	180	309	H	1	170	297	H	0	106	383	H	3	97	368	H	1	207	917	H	3	101	364	H
MLP	371	206	0	R	405	153	1	R	367	119	3	R	433	126	2	R	904	216	5	R	433	125	1	R
	158	880	87	M	131	701	74	M	117	286	86	M	180	588	138	M	290	629	206	M	198	624	84	M
	0	211	278	H	2	191	275	H	5	117	367	H	3	125	340	H	6	240	879	H	4	150	314	H
C4.5	346	228	3	R	379	181	1	R	371	106	12	R	438	110	13	R	445	108	8	R	445	108	8	R
	221	775	129	M	157	687	62	M	156	201	132	M	251	430	225	M	268	517	121	M	268	517	121	M
	6	205	278	H	6	215	247	H	18	113	358	H	11	104	353	H	9	172	287	H	9	172	287	H
LMT	329	247	1	R	377	183	1	R	387	99	3	R	469	90	2	R	918	191	16	R	468	91	2	R
	110	905	110	M	104	746	56	M	126	259	104	M	237	487	182	M	270	648	207	M	228	568	110	M
	0	179	310	H	1	178	289	H	5	96	388	H	5	93	370	H	14	193	918	H	7	131	330	H
RF	345	232	0	R	383	177	1	R	380	107	2	R	458	97	6	R	951	164	10	R	421	139	1	R
	142	868	115	M	131	717	58	M	129	252	108	M	250	478	178	M	201	764	160	M	180	634	92	M
	3	200	286	H	5	179	284	H	7	115	367	H	7	95	366	H	14	149	962	H	7	151	310	H

males; Tested TD, RUS and SMOTE models showed differences of 0.06, 0.07 and 0.07 respectively between the highest and the lowest performing algorithms. C4.5 showed consistency in achieving the least TPR among all tested models.

At a class level, taking the risk group into account for this comparison, Fig. 10 demonstrates relatively large differences between the minimum and the maximum TPR test performances for R class in each cohort.

In the females, tested TD, RUS and SMOTE models showed a high R class accuracy differences of 0.34, 0.13 and 0.25 respectively between the highest and the lowest performing algorithms. A lesser TPR differences were found in the males TD, RUS and SMOTE models of 0.08, 0.13 and 0.12 respectively between the highest and the lowest performing algorithms. NB showed consistency in scoring the lowest TPR among all tested models.

Fig. 9 Comparison of performance metrics across trained models in different cohorts



4.3 Attribute information gain results

In the female cohort training datasets, when considering the IG for each variable across all datasets (Fig. 12), the IG increased in each attribute for RUS and SMOTE datasets compared to the TD. By comparing the IG ranking of variables in each dataset, it is apparent that WC achieved the highest IG value in all the three datasets. The dominance in WC ranking was also accompanied by an increase of its values (from TD to RUS and SMOTE). Such an increase

correlates directly with the increase in class R TPR performance in all trained models except for NB where RUS model overtook SMOTE by a small TPR positive margin of 0.092. From Fig. 12, SMOTE boosted the information within each variable (Table 5). This boost, in turn, increased the ability to differentiate class R from other classes in the TD, which in turn increases the class R TPR (see Fig. 10). The APA multi-dimensional visualisation (see Fig. 13) shows the improved class R discrimination per dataset.

Fig. 10 Risk class TPR performance for trained and tested models per cohort

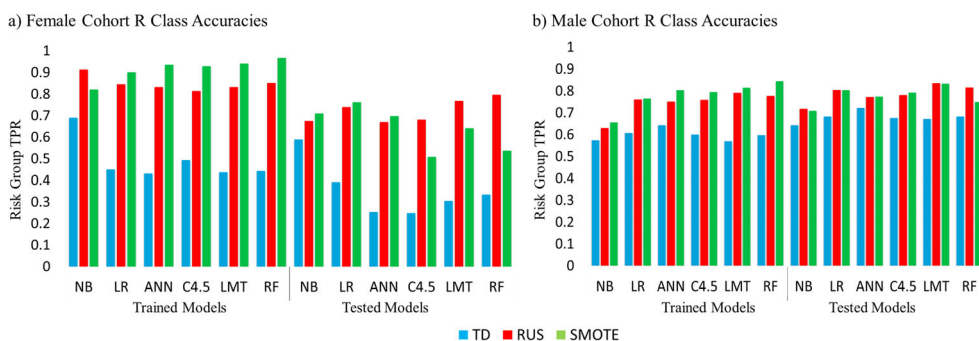
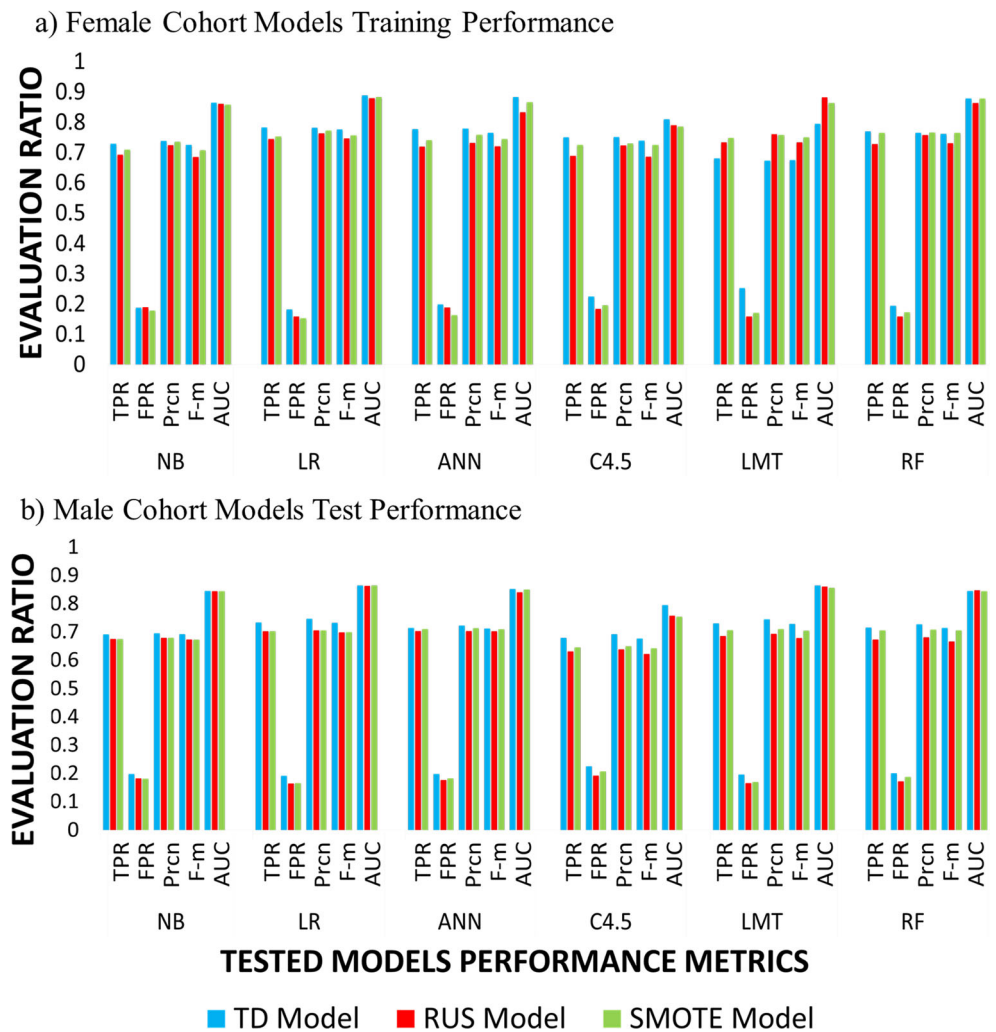


Fig. 11 Comparison of performance metrics across all tested models



In the male cohort training datasets, when considering the measured IG for each variable across all datasets (Fig. 12), it is observed that the IG increased in each attribute for SMOTE dataset and some of the attributes for RUS dataset compared to the TD. By comparing the IG ranking of variables in each dataset, it was apparent that waist circumference (WC) achieved the highest IG value in all the TD and RUS datasets while BMI achieved the highest IG value in the SMOTE dataset. The

advancement in BMI ranking in SMOTE dataset correlates directly with the increase in class R TPR performance in all trained models. SMOTE resampling technique amplified the information within each variable (Fig. 12). This amplification, in turn, increased the class R border density with other classes in the training dataset, which in turn increased class R TPR in training (see Fig. 10). The APA visualisation showing the enhancement in class R borders density per dataset is shown in Fig. 13.

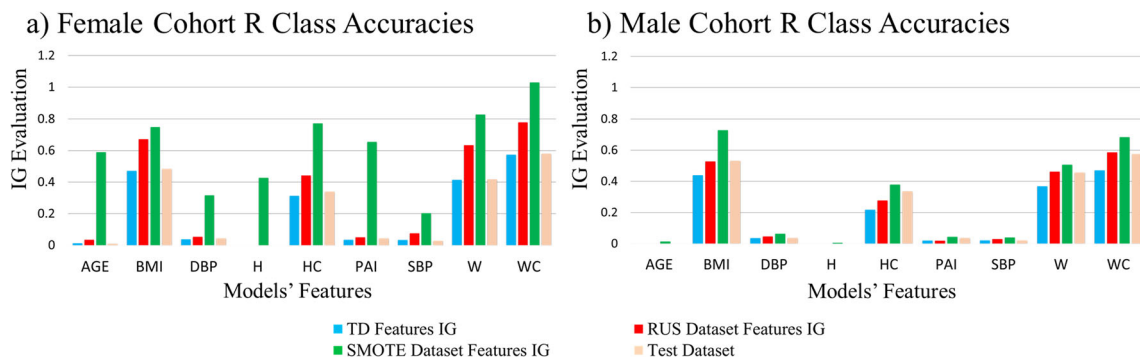


Fig. 12 Information Gain evaluation comparison of all variables per dataset

Table 5 The Information Gain evaluation of all features per dataset

Dataset Features	Female Cohort Dataset				Male Cohort Dataset			
	TD	RUS	SMOTE	Test Dataset	TD	RUS	SMOTE	Test Dataset
Age	0.0125	0.0346	0.5900	0.0092	0.0000	0.0000	0.0138	0.0000
BMI	0.4717	0.6725	0.7488	0.4839	0.4392	0.5279	0.7287	0.5315
DBP	0.0375	0.0537	0.3163	0.0435	0.0360	0.0460	0.0637	0.0364
H	0.0000	0.0000	0.4266	0.0000	0.0000	0.0000	0.0060	0.0000
HC	0.3130	0.4420	0.7725	0.3399	0.2182	0.2766	0.3790	0.3361
PAI	0.0346	0.0503	0.6546	0.0435	0.0204	0.0194	0.0445	0.0363
SBP	0.0331	0.0755	0.2034	0.0278	0.0211	0.0292	0.0399	0.0213
W	0.4148	0.6341	0.8284	0.4172	0.3685	0.4624	0.5071	0.4569
WC	0.7543	0.7781	0.0284	0.5806	0.4700	0.5857	0.6835	0.5751

4.4 Domain experts’ results

The misclassification of healthy subjects by a predictive model could result in costly and unnecessary follow-up examinations whilst false-negative misclassifications might result in an individual not receiving an important intervention. In this application, apart from potential cost, there would be few adverse effects associated with healthy/moderate risk subjects being misclassified, as such subjects would be encouraged to undertake lifestyle-based interventions to improve their health. Therefore, in this scenario the best models to adopt would be those which minimise the number of subjects misclassified as at ‘risk’, so they may initiate interventions at an appropriate time. Confusion matrices play an essential role in helping researchers define the best-suited model for use in future trials. When analysing the confusion matrices (Table 4), from both the female and male cohorts three models from each cohort were identified as satisfying the domain experts’ criteria. These models are reported in Table 6. They may not

necessarily occupy the highest ranks when their performance metrics were compared to the others.

5 Discussion

The overall goal of this study was to predict visceral adipose tissue (VAT) content in male and female participants from the UKBB and to apply machine learning methods to classify these subjects into risk categories. VAT has consistently been shown to be associated with the development of metabolic conditions such as coronary heart disease and type-2 diabetes. The ability to predict and classify this variable, using simple anthropometry without the need for costly MRI scanning, will have a significant impact on the identification of subjects likely to benefit most from life-style based interventions [41]. The models tested here input features that include age, waist and hip circumferences, weight, height, BMI, blood pressure and

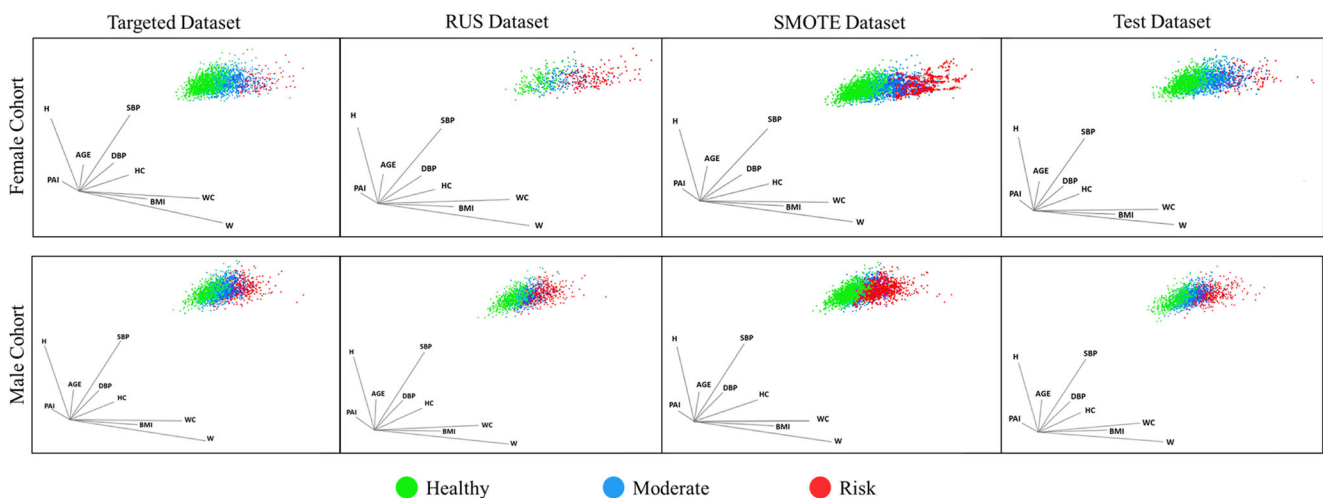


Fig. 13 Adaptive projection visualisation of all classes and the effect of sampling methods

Table 6. Domain compliant prediction models

a Selected Female Cohort VAT Prediction Models

	Model		Dataset	
	Algorithm			
	LMT	LR	RF	
	RUS	SMOTE	RUS	
Performance	Metrics			
	CCI	TPR	FPR	Pern
Performance	Training			
	CCI	TPR	FPR	Pern
Performance	Test			
	CCI	TPR	FPR	Pern
Confusion Matrices	Trained			
	Actual			
Confusion Matrices	Tested			
	Actual			

b Selected Male Cohort VAT Prediction Models

	Model		Dataset	
	Algorithm			
	LMT	LMT	RF	
	RUS	SMOTE	RUS	
Performance	Metrics			
	CCI	TPR	FPR	Pern
Performance	Training			
	CCI	TPR	FPR	Pern
Performance	Test			
	CCI	TPR	FPR	Pern
Confusion Matrices	Trained			
	Actual			
Confusion Matrices	Tested			
	Actual			

a) For females n (LMT RUS Trained) = 486; n (LR SMOTE Trained) = 3426; n (RF RUS Trained) = 486. n (all Tested) = 2035
 b) For males n (LMT RUS Trained) = 1467; n (LMT SMOTE Trained) = 3375; n (RF RUS Trained) = 1467. n (all Tested) = 1935. For the F-m metric, m = 1

level of physical activity, all variables previously demonstrated to significantly correlate with VAT [42].

Previous UKBB studies [39, 40] have demonstrated significant correlations of anthropometry measurement and physical activities with VAT, with significant gender differences in the distribution of VAT, as well as by age; Hence separate models were built for females and males

participants. In the same study, an index of physical activity, the UKBB-PAI, was proposed which correlated more strongly with VAT outcomes than established questionnaires, such as the International physical activity questionnaire (IPAQ) and lifestyle Index. Additionally, its findings challenged previous studies [42–44], and describes only a weak correlation between age with VAT, even after

adjusting for BMI, and UKBB-PAI. It was also noted that the influence of UKBB-PAI parameters was comparable to that of age, and that it provided more effective means representing the physical activity measures to discriminate between Health, Moderate and Risk classes.

With domain experts' advice, the current study selected the above-mentioned variables (age, blood pressure, body mass index, height, hip circumference, physical activity index, waist circumference and weight) as input features on which to base the ML VAT prediction models. However, to understand the influence of each feature on VAT and their reliability to predict three distinct ranges associated with various long-term conditions, information theory was used to evaluate each feature in relation to the 3 different classes, Healthy, Moderate and at Risk. The IG evaluation algorithm was utilised to evaluate the worth of each input feature independently [15] against the class unlike correlation analysis carried out in previous studies [39]. One way to interpret the calculated IG values is the possible presence of associations between each feature and the class labels in each cohort. In the female cohort, the strength of the association in the IG varies (see Table 5), with HC, WC, BMI and W providing the greatest contribution, whilst the physical activity, age, H, SBP and DBP showed the least in both TD and RUS training datasets. A similar dominance in IG ranking is observed in the SMOTE dataset, with HC, WC, BMI and W showing the strongest associations with VAT. An analogous pattern was found in male subjects. This information theory approach into the models' features adds an additional layer of details to observed correlations reported in previous studies by describing the strength of each feature to discriminate between the Health, Moderate and at Risk classes.

When considering the TD and RUS datasets which contain observations from the participants rather than generated artificial synthetic data, there was no association between age and VAT (given zero IG) in the males and only a weak association in the female cohort, with the discretised VAT ranges similar to the weak correlation described in previous studies. Though this may challenge previous studies that have reported a linear relationship between age and VAT [42–44], our results may reflect the somewhat smaller age range included in the UKBB (40–70 yrs), compared with previous studies (17–70 yrs) [42]. However, this may also relate to a data problem in the machine learning community known as data heterogeneity [45].

The lack of association of UKBB-PAI with discretised VAT classes, reflects the previously reported [46, 47] low correlation between physical activity and this fat depot, and may in part arise from to the poor reliability of the recorded frequencies and durations of physical activities. The level of granularity in the input data variables always determines the level of details in the prediction model possible outputs.

Depending on the assessment design, detailed observations may be grouped during or after data collection into frequencies, categories and scores. This grouping is considered a variable transformation. Variable transformation aims to create better features at exposing patterns in the data. However, the transformation process could also lead to engineering a new feature that is less powerful suppressing important trends offered by its detailed (raw) components.

It could also be argued that the implementation of such low-cost measures may lack the susceptibility to errors if studied within large populations [48]. However, there are many newly developed physical activities questionnaires (PAQs) which do not appear to perform substantially better than existing tests with regard to reliability and validity [49–51]. The variability of these PAQs and their ineffectiveness leads to a cause known in the data science community as *detail aggregation*. Variables in datasets often fall within two types; either detailed (Granular) or aggregated (Summaries). ML modeling prefer detailed variables over summary variables. Detailed data often represent summary variables and better at showing patterns. Take daily walking which forms part of PAI calculations as an example. Previous studies [52, 53] showed that daily walking is linked to reductions in VAT. However, its significance is curbed when combined with other variables in PAI calculations. Data granularity is a macro structural feature. Granularity refers to the amount of detail captured in any measurement such as time to the nearest minute, the nearest hour, or simply differentiating morning, afternoon, and night, for instance. Decisions about macro structure have an essential impact on the amount of information that a data set carries, which, in turn, has a very significant effect on the resolution of any model built using that data set [54]. Therefore, we must acknowledge that physical activity is a complex behavior that is hard to measure accurately even at a low degree, in case of memory recollection, or a high degree, by using monitoring electronic devices. However, it is a real challenge to record the interactions among physical activity various elements (variables). The PAI structure that combines sets of variables with transformed scores could be introducing bias, which stresses the natural structure of the original variables states in a dataset so that the data is distorted. Hence the PAI may be less representative of the real world than the original, unbiased variables form.

The understanding of the effect of data aggregation by domain experts enhances feature selection strategies of how variables are used in predictive modeling. Some derived (aggregated) variables may increase the representation of trend within a dataset which by turn, show higher IG evaluation and act as a stronger predictor in modeling. For example, BMI is directly obtained from height and weight (calculated as weight in kilo-grams (W) divided by height (H) in meters squared). From our analysis, H maintained its IG evaluation to zero in both RUS and TD datasets, by dividing body mass

over two exponents of the base H , this seems to expose better trends. Aggregated variables may require checking for calculations integrity from detailed variables. For numeric features, aggregated variables come in many forms such as averages, sums, multiplication and ratios. Categorical features can be combined into a single feature containing combination of different categories. Variable aggregation must not be overdone as to not overfit models due to misleading combined features. Wrongly derived variables may show false significance or insignificance in the analysis [54].

For ML modeling, tackling the imbalanced class problem has a significant impact on the performance of standard machine learning algorithms. Classification performance in the training phase is severely impacted by class separability. Training standard ML algorithms with highly imbalanced overlapping classes without any adjustment to the training set results in an accuracy bias towards the majority class. In this study, we observed that applying the two methods (RUS and SMOTE) was used to adjust the class imbalance in the classification training phase at the dataset level, which in turn, amplified the IG in many input features. It remains unclear as to whether other remedies for imbalanced data classifications, such as Cost-Sensitive and Ensembles Learning (which are implemented at algorithmic level), could result in better performances [4, 6, 55]. The advantages of sampling techniques evaluated here, however, include simplicity and transportability. Nevertheless, they are limited by the amount of IG manipulation as a result of their application resulting in biased predictions towards the minority class. The excessive use of such techniques could result in overfitting of the models.

In this study, for the female cohort case, the original dataset was highly imbalanced. Traditional ML algorithms were sensitive to higher information gains. They tended to produce superb performance results in training, but when testing the models, the overall model accuracy often dropped below the training phase performance.

However, for the male cohort, the class imbalance in the original dataset was less severe; therefore, traditional ML algorithms were less sensitive to higher information gains and tended to produce close performance results in training and test. The overall model accuracy often dropped below the training phase performance, which was the case for all models trained with the SMOTE dataset. On the contrary, the models' test accuracy outperformed the training accuracy when each algorithm was trained on TD; this situation also occurred in NB, LR and MLP trained with RUS dataset. The cause of such competitive accuracy test results may be attributed to the increase in IG per feature in the test dataset as compared to the TD (Fig. 12). A higher IG in a variable indicates higher observations' purity per class. Having higher IG in multiple features enhances class separability and leads to improvement in classification accuracy. In other words, the higher the IG in a dataset the easier the dataset to be learned and to be predicted.

In both cohorts, the UKBB datasets utilised in this study showed that applying the correct level of sampling without disrupting the original data distribution, together with the desired choice of performance metrics and slight manipulation of IG levels produced a prediction solution which could be developed further with algorithmic modifications [8]. Among all eighteen models for each cohort presented in this study, six models satisfied the domain experts' success criteria for this specific domain problem. For the female cohort, these were LMT and RF built with RUS sampled dataset, and LR built with SMOTE sampled dataset. For the male cohort, they were LMT and RF built with RUS sampled dataset, and LMT built with SMOTE sampled dataset.

The difference in ML algorithms learning schemes proved to have a minimal impact on the whole model accuracy. ML algorithms are biased towards achieving the highest model's accuracy. But the effect of learning scheme becomes largely noticeable in imbalanced datasets when the minority classes accuracies are compared. In the testing results analysis, learning schemes impact was seen to increase with the class imbalance severity in datasets compared to balanced datasets.

This domain problem is the first to use the discretised MRI VAT variable ranges to describe the health status of participants and to label instances. At present, it would be impractical to compare the results of this study to any other research from the same domain. However, this work will be followed by further analyses where additional methods to improve the outcomes will be investigated.

6 Conclusion

Our study shows that the application of traditional machine learning algorithms to datasets of phenotype variables offers a fast and inexpensive solution to predict visceral fat by aligning the classification task to predict specific VAT ranges. The selection of a multi-class prediction task in this study is strategic. It identifies individuals who are at higher risk of developing metabolic conditions and are more likely to benefit from focused lifestyle intervention to reduce visceral fat. The design of the case study of a multi-class prediction, by separating the risk group from a moderate group, helped in selecting models that minimise incorrect classification of those who are at high risk as healthy. Achieving a zero False Negative Rate (FNR) when classifying risk patients as healthy guarantees that any individual to miss treatment intervention belongs to the moderate group rather than the risk group. The process of training various ML algorithms with 10-Fold Cross-Validation and testing the models with external groups of females and males of similar ratio to the training data makes this study suitable for follow-up research in medical screening to identify subjects that may require treatment intervention.

Acknowledgements This work was funded by the Quintin Hogg Trust PhD Scholarship Awards, supported by experts from the Health and Innovation Ecosystem at the University of Westminster and the UK Biobank.

Compliance with ethical standards

Conflict of interest The authors whose names are listed immediately below the manuscript title certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Ethics approval The questionnaire and methodology for this study was approved by the UK Biobank Ethics Committee. The UKBB had approval from the North West Multi-Centre Research Ethics Committee (MREC), and written consent was obtained from all participants before their involvement. The data was acquired through the UK Biobank Access Application number 23889.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Yang Q, Wu X. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*. 2006;5(4):597–604.
2. Gu J, Zhou Y, Zuo X. Making Class Bias Useful: A strategy of learning from imbalanced data. In: Yin H, Tino P, Corchado E, Byrne W, Yao X, editors. *IDEAL 2007, LNCS*, vol. 4881. Heidelberg: Springer; 2007. p. 287–95.
3. More A. Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv:1608.06048 [stat. AP]* (2016).
4. Weiss GM, McCarthy K, Zabar B. Cost-Sensitive Learning vs Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs? In: *Proceedings of the 2007 International Conference on Data Mining*, pp. 35–41, Las Vegas, USA (2007).
5. Bekkar M, Taklit AA. Imbalanced data learning approaches review. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*. 2013;3(4):15–33.
6. Ensemble Learning to Improve Machine Learning Results, <https://blog.statsbot.co/ensemble-learning-d1dcd548e936>, last accessed: 2019/02/19.
7. Dzeroski S, Zenko B. Is combining classifiers better than selecting the best one? In: *Proceedings of the Nineteenth International Conference on Machine Learning*, San Francisco, Morgan Kaufmann (2002).
8. Choi JM. A Selective Sampling Method for Imbalanced Data Learning on Support Vector Machines. Iowa State University (Graduate Theses and Dissertation) (2010).
9. Unbalanced Data Is a Problem? No, Balanced Data Is Worse, <https://matloff.wordpress.com/2015/09/29/unbalanced-data-is-a-problem-no-balanced-data-is-worse/>, last accessed: 2019/02/24.
10. When should I balance classes in a training data set? <https://stats.stackexchange.com/questions/227088/when-should-i-balance-classes-in-a-training-data-set>, last accessed: 2018/11/22.
11. Bharat RR, Fung G, Rosales R. On the Dangers of Cross-Validation. An Experimental Evaluation. In: *Proceedings of the 2008 SIAM International Conference on Data Mining*, pp. 588–596 (2008).
12. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–57.
13. Faith J, Mintram R, Angelova M. Gene expression targeted projection pursuit for visualising gene expression data classifications. *Bioinformatics*. 2006;22(21):2667–73.
14. Information Gain Which test is more informative? <https://homes.cs.washington.edu/~shapiro/EE596/notes/InfoGain.pdf>, last accessed 2019/03/29.
15. Quinlan JR. Induction of decision trees. *Mach Learn*. 1986;1(1): 81–106.
16. Wang YC, McPherson K, Marsh T, Gortmaker SL, Brown M. Health and economic burden of the projected obesity trends in the USA and the UK. *Lancet*. 2011;378(9793):815–25.
17. Sam S, Mazzone T. Adipose tissue changes in obesity and the impact on metabolic function. *Transl Res*. 2014;164(4):284–92.
18. Dattilo AM, Kris-Etherton PM. Effects of weight reduction on blood lipids and lipoproteins: a meta-analysis. *Am J Clin Nutr*. 1992;56(2):320–8.
19. Fox CS, Massaro JM, Hoffmann U, Pou KM, Maurovich-Horvat P, Liu CY, et al. Abdominal visceral and subcutaneous adipose tissue compartments. *Circulation*. 2007;116(1):39–48.
20. Després JP, Lemieux I, Bergeron J, Pibarot P, Mathieu P, Larose E, et al. Abdominal obesity and the metabolic syndrome: contribution to global Cardiometabolic risk. *Arterioscler Thromb Vasc Biol*. 2008;28(6):1039–49.
21. Chin SH, Kahathuduwa CN, Binks M. Physical activity and obesity: what we know and what we need to know*. *Obes Rev*. 2016;17(12):1226–44.
22. Golabi P, Bush H, Younossi ZM. Treatment strategies for nonalcoholic fatty liver disease and nonalcoholic Steatohepatitis. *Clinics in Liver Disease*. 2017;21(4):739–53.
23. Uusitupa M, Lindi V, Louheranta A, Salopuro T, Lindström J, Tuomilehto J. Long-term improvement in insulin sensitivity by changing lifestyles of people with impaired glucose tolerance. *Diabetes*. 2003;52(10):2532–8.
24. Brouwers B, Hesselink MKC, Schrauwen P, Schrauwen-Hinderling VB. Effects of exercise training on intrahepatic lipid content in humans. *Diabetologia*. 2016;59(10):2068–79.
25. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):e1001779.
26. Information gain, mutual information and related measures - Cross Validated, <https://stats.stackexchange.com/questions/13389/information-gain-mutual-information-and-related-measures>, last accessed 2018/10/22.
27. Haddow C, Perry J, Durrant M, Faith J. Predicting functional residues of protein sequence alignments as a feature selection task. *International Journal of Data Mining and Bioinformatics*. 2011;5(6):691–705.

28. Drummond C, Holte RC. C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In: Proceedings of the International Conference on Machine Learning, Workshop Learning from Imbalanced Data Sets II (2003).
29. Manning C, Raghavan P, Schütze H. Introduction to information retrieval. *Nat Lang Eng.* 2010;16(1):100–3.
30. Zhang H. The optimality of naive Bayes. *American Association for Artificial Intelligence* (2004).
31. Landwehr N, Hall M, Frank E. Logistic model trees. *Mach Learn.* 2005;59(1–2):161–205.
32. Ayer T, Chhatwal F, Alagoz O, Kahn CE, Woods RW, Burnside ES. Comparison of logistic regression and artificial neural network models in breast Cancer risk estimation. *Radio Graphics.* 2010;30(1):13–22.
33. Quinlan JR. Improved use of continuous attributes in C4.5. *J Artif Intell Res.* 1996;4:77–90.
34. Witten IH, Frank E. *Data Mining, Practical Machine Learning Tools and Techniques.* 2nd edn. Elsevier Inc (2005).
35. Witten IH, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, San Francisco, 2000.
36. Jonsdottir T, Hvanngberg ET, Sigurdsson H, Sigurdsson S. The feasibility of constructing a predictive outcome model for breast cancer using the tools of data mining. *Expert Syst Appl.* 2008;34(1):108–18.
37. Maheshwari S, Agrawal J, Sharma S. A new approach for classification of highly imbalanced data sets using evolutionary algorithms. *International Journal of Scientific & Engineering Research.* 2011;2(7):1–5.
38. Computing Precision and Recall for Multi-Class Classification Problems, <http://text-analytics101.rxnlp.com/2014/10/computing-precision-and-recall-for.html>, last accessed 2018/08/02.
39. Parkinson JR, et al. Visceral adipose tissue, thigh adiposity and liver fat fraction: a cross-sectional analysis of the UK biobank. *UK Biobank* (2019).
40. Parkinson, JR, Gerbault P, Alenaini W, Elliot B, Wilman H, Bell JD, Thomas EL. Physical activity, visceral adipose tissue, thigh adiposity and liver fat fraction: a cross sectional analysis of the UK biobank. *The UK biobank* (2019). Submitted.
41. Shuster A, Patlas M, Pinthus J, Mourtzakis M. The clinical importance of visceral adiposity: a critical review of methods for visceral adipose tissue analysis. *Br J Radiol.* 2012;85(1009):1–10.
42. Thomas EL, Parkinson JR, Frost GS, Goldstone AP, Dore CJ, McCarthy JP, et al. The missing risk: MRI and MRS phenotyping of abdominal adiposity and ectopic fat. *Obesity (Silver Spring).* 2012;20(1):76–87.
43. Palmer BF, Clegg DJ. The sexual dimorphism of obesity. *Mol Cell Endocrinol.* 2015;402:113–9 Epub 2015/01/13.
44. Machann J, Thamer C, Schnoedt B, Haap M, Haring HU, Claussen CD, et al. Standard-ized assessment of whole body adipose tissue topography by MRI. *J Magn Reson Imaging.* 2005;21(4):455–62.
45. Bagging and Random Forest Ensemble Algorithms for Machine Learning, <https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/>, last accessed: 2018/10/22.
46. Bisschop CN, Peeters PH, Monninkhof EM, Van Der Schouw YT, May AM. Associations of visceral fat, physical activity and muscle strength with the metabolic syndrome. *Maturitas.* 2013;76(2):139–45.
47. Pasdar Y, Darbandi M, Mirtaher E, Rezaeian S, Najafi F, Hamzeh B. Associations between muscle strength with different measures of obesity and lipid profiles in men and women: results from RaNCD cohort study. *Clin Nutr Res.* 2019;8(2):148–58 Epub 2019 Apr 26.
48. Van Poppel MN, Chinapaw MJ, Mokkink LB, van Mechelen W, Terwee CB. Physical activity questionnaires for adults: a systematic review of measurement properties. *Sports medicine (Auckland, NZ).* 2010;40(7):565–600. Epub 2010/06/16.
49. Helmerhorst HJ, Brage S, Warren J, Besson H, Ekelund U. A systematic review of reliability and objective criterion-related validity of physical activity questionnaires. *The international journal of behavioral nutrition and physical activity.* 2012;9:103 Epub 2012/09/04.
50. Hagstromer M, Bergman P, De BI, Ortega FB, Ruiz JR, Manios Y, et al. Concurrent validity of a modified version of the International Physical Activity Questionnaire (IPAQ-A) in European adolescents: The HELENA Study. *Int J Obes (Lond).* 2008;32(Suppl 5):S42–S8.
51. Ferrari P, Friedenreich C, Matthews CE. The role of measurement error in estimating levels of physical activity. *Am J Epidemiol.* 2007;166(7):832–40.
52. Miyatake N, Nishikawa H, Morishita A, Kunitomi M, Wada J, Suzuki H, et al. Daily walking reduces visceral adipose tissue areas and improves insulin resistance in Japanese obese subjects. *Diabetes Res Clin Pract.* 2002;58(2):101–7 Epub 2002/09/06.
53. Mytton OT, Ogilvie D, Griffin S, Brage S, Wareham N, Panter J. Associations of active commuting with body fat and visceral adipose tissue: a cross-sectional population based study in the UK. *Prev Med* 2017. Epub 2017/10/17.
54. Pyle D. *Data preparation for data mining.* 1st edn. Morgan Kaufmann Publishers, Inc (1999).
55. Grainger AT, Tustison NJ, et al. Deep learning-based quantification of abdominal fat on magnetic resonance images. *PLOS ONE* (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.