Machine Learning Classification of Females Susceptibility to Visceral Fat Associated Diseases

M. Aldraimli [1], D. Soria [1], J. Parkinson [2], B. Whitcher [2], E. L. Thomas [2], J. D. Bell [2], T. J. Chaussalet [1] and M. V. Dwek [2]

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

[2] School of Life Sciences, University of Westminster, London W1W 6UW, UK

m.aldraimli@my.westminster.ac.uk
d.soria@westminster.ac.uk

Abstract. The problem of 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 these algorithms is to predict dichotomous responses (e.g. healthy/at risk) based on several features. Similarly to statistical inference models, also ML models are subject to the common problem of class imbalance. Therefore, they are affected by the majority class increasing the false negative rate.

In this paper, we built and evaluated eighteen ML models classifying approximately 4300 female participants from the UK Biobank into three categorical risk statuses based on responses for the discretised visceral adipose tissue values from magnetic resonance imaging. We also examined the effect of sampling techniques on classification modelling when dealing with class imbalance.

Results showed that the use of sampling techniques had a significant impact. They not only drove an improvement in predicting patients risk status, but also facilitated an increase in the information contained within each variable. Based on domain experts criteria, the three best models for classification were finally identified.

These encouraging results will guide further developments of classification models for predicting visceral adipose tissue without the need for a costly scan.

Keywords: Supervised Learning, Imbalanced Data, UK Biobank, Random Under-Sampling, Synthetic Minority Over-Sampling Technique, Visceral Adipose Tissue.

1 Introduction

Real-world data are often imbalanced and lack uniformly distribution across classes. Classification of imbalanced datasets is one of the challenges across several industrial and research domains [1]. There are multiple approaches to tackle class imbalance [2], simplest approaches include, for example, Data Enrichment. Others more sophisticated methods include various sampling techniques [3], cost-sensitive learning [4], [5] and feature selection; or more complex strategies including meta-learning [6], combining classifiers [7], and algorithmic modifications [8].

When sampling methods are applied, questions over their suitability are often raised [9]. For example: is the new re-sampled 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 also compromise accuracy measures by deceiving the process of cross-validations sampling [11].

In this paper we compared the classification performance of six ML algorithms (Naïve Bayes, Logistic Regression, Artificial Neural Network, Decision Tree, Logistic Model Tree, and Random Forest) when using RUS [8] and SMOTE [12] sampling techniques on highly imbalanced data to predict visceral adipose tissue (VAT) disease risk in a multi-class classification problem, and to suggest the most suitable models to meet 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 Attribute Evaluation (IG) [14], [15].

The paper is structured as follows: in Section 2, the domain problem and all the methods and approaches used in this study are presented. Then in Section 3, the experiments' results are introduced, while Section 4 is reserved for discussion and conclusions.

2 Materials and Methods

2.1 The Domain Problem

Obesity affects an increasing number of adults in the UK [16], with obesity-associated changes in adipose tissue (AT) predisposing to metabolic dysregulation [17]. Distribution of AT, in particular the accumulation of VAT and liver fat, are a key factor in determining susceptibility to disease [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], aiming to reduce VAT and liver fat, via exercise, 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 requisite imaging techniques. The UK Biobank provides a comprehensive means of assessing the relationship between body composition and lifestyle in a large population-based cohort of adults aged 40-70 years, recruited between 2007 and 2010 [25]. The primary goal of this study was to identify the best model able to predict VAT levels in a cohort of female individuals from the UK Biobank. The UK Biobank had approval from the North West Multi-Centre Research Ethics Committee (MREC) and written consent was obtained from all participants prior to their involvement. The data was acquired through the UK Biobank Access Application number 23889. The study was a cross-sectional assessment of 4327 female individuals from the UK Biobank multimodal imaging cohort [25]; aged 40-73 years and scanned chronologically between August 2014 and September 2016. The analysis of male subjects VAT is outside the scope of this paper.

2.2 Methodology

Multi-class classification ML models were applied with the aim of predicting susceptibility to disease (risk) based on the discretised amount of VAT. A subset of 2292 subjects was randomly selected from the original 4327 females and used to train six ML algorithms using 10-fold cross-validation in three different scenarios. The models were tested on the remaining 2035 cases. Fig. 1 shows the methodology: multiple imbalanced datasets with the same predictor variables were modified with sampling techniques, and used for modelling using the six ML algorithms. The accuracies of the models were compared after the training phase. IG was monitored for all predictor variables at every stage.

Information Gain Evaluation Algorithm (IG). 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 which ranks the attributes by their individual influence on the class [14], [15], [26].

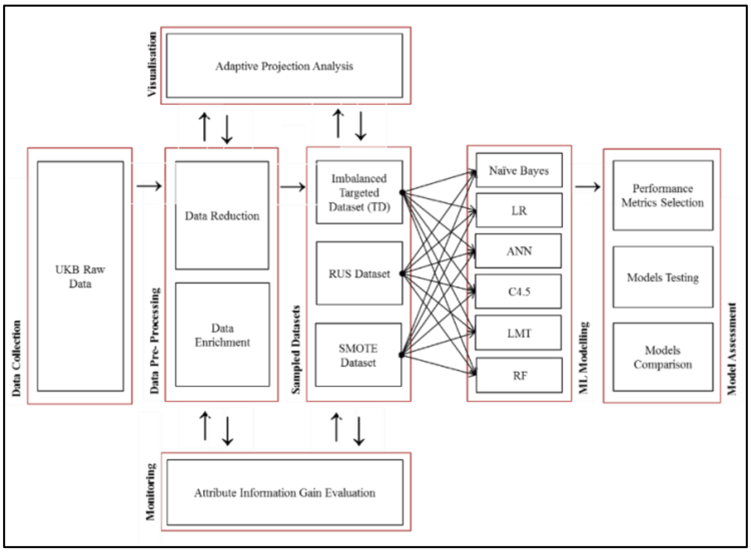


Fig. 1. Methodology adopted in this work, showing the different steps followed.

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].

2.3 Data Preparation

Random Undersampling (RUS). This approach consisted of selecting a subset of the majority class to balance the data [8]. In this approach 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 value of *r* = 0.5.

$$r = \frac{|S|}{|L|} = 0.5$$

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. 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, artificial observations were generated and spread in the area between *x* and 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.

2.4 ML Classification Algorithms

Naïve Bayes (NB). A probabilistic 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 have been reported to be effective in complex real-world situations [30].

Logistic Regression (LR). LR is a deterministic technique which produces a probability-based model that takes into account 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, especially high dimensional datasets characterised by noise and missingness (pollution). Neural networks comprise elementary autonomous computational units, known as neurons. Neurons are inter-connected via weighted connections and are organised in layers (an input layer, hidden layers and an output layer). In this study, we used a Multilayer Perceptron ANN with a sigmoid activation function [17].

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 a model with a tree structure but with LR functions at the leaves level. The LMT structure comprises a set of non-terminal nodes and a set of leaves (terminal nodes). LMT has been designed to adapt to small data subsets where a simple linear model offers best bias-variance trade-off [31].

Random Forest (RF). RF is a generalisation of standard decision trees proposed by Brieman 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), with the C4.5 using the J48 implementation.

2.5 Model Evaluation

In agreement with domain experts, we chose several measures to evaluate the performance of each model. These measures included accuracy (later reported as 'CCI%') true positive rate (also known as sensitivity or recall, 'TPR'), specificity, false positive rate ('FPR'), precision ('Prcn'), area under the receiver operator 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].

2.6 Experimental Design

The analysis was performed to predict VAT related disease susceptibility based on discretised MRI response labels: Healthy, Moderate and Risk defined according to VAT volume. If VAT volume was ≤2 litres then the subject was deemed 'Healthy' (H). If VAT volume was >2 litres but ≤5 litres, then the class was 'Moderate' (M). If VAT volume was >5 litres, then the patient was at ‘Risk’ (R) [39]. The training datasets contained ten data variables reported in Table 1, with the VAT in litres being the class determination response variable.

Table 1. Summary statistics of TD variables

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)
Predictors 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 m)	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)

Targeted dataset (TD). The TD was the first dataset we considered for modelling. The TD contained 2292 female records, from the UK Biobank cohort [45]. Table 1 shows the summary statistics of all TD’s variables. The TD was highly imbalanced: class H had 1002 subjects, the M class had 1128 subjects, and the minority R class contained only 162 subjects.

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 RUS to the TD, each of the H, M and R classes ended up with 162 subjects.

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. The H class had 1002 subjects, the M class had 1128 subjects and the R class contained 1296 subjects. The effect of SMOTE can be observed via APA visualisation in Fig. 2.

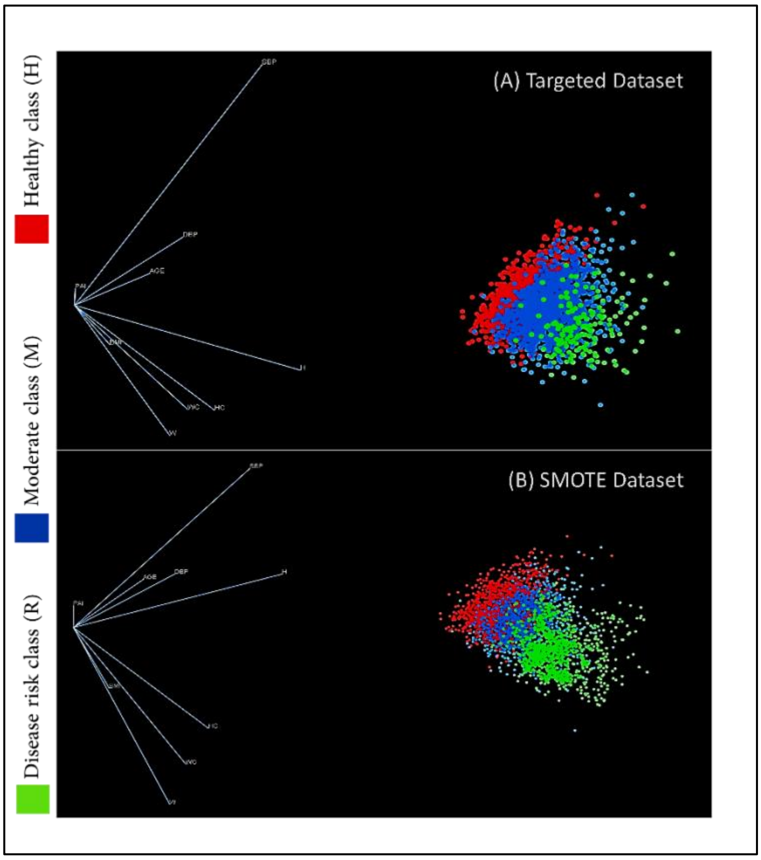


Fig. 2. APA visualisation of SMOTE dataset variables

We used the IG Evaluator Algorithm 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 will be presented in Section 3.

The Test Dataset. The ML models were tested on the remaining 2035 individuals from the original 4327 UK Biobank cohort. The same ten variables as per the training datasets were available. Table 2 shows their summary statistics. Similarly to TD, the Test Dataset was also highly imbalanced: class H had 823 subjects, the M class had 1039, and class R contained only 173 subjects.

Table 2. Summary statistics of test set variables

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 m)	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)

3 Results

3.1 Models Training Results

From Table 3, the model training accuracies (CCI%) of all methods were computed and they showed that resampling methods resulted in an improvement in CCI% as compared to the original TD. Fig. 3 shows that LR, ANN, C4.5 and RF models training performances using the RUS dataset worsened compared to the same algorithms trained on TD. The ROC for each of the trained models ranged between 0.783 (for RF on SMOTE) and 0.96 (for C4.5 on TD). These values indicate that the trained models did not sacrifice a lot of precision of the system to get a good recall on the observed data points. The RF model achieved the highest TPR (0.850) when trained on the SMOTE dataset, whilst the C4.5 model achieved the lowest TPR (0.714) when trained on the RUS dataset.

By observing the confusion matrices for all models after training on all the TD and RUS datasets, and bearing in mind that all comprise the same risk group participants, it is clear that the number of incorrectly classified instances for the R class significantly decreased for the models trained on the RUS dataset compared to those trained on the original TD. However, when evaluating the minority class accuracy performance in Fig. 4, it is notable that all trained models benefitted from the sampling methods, exhibiting significant TPR improvement for the R group in each model.

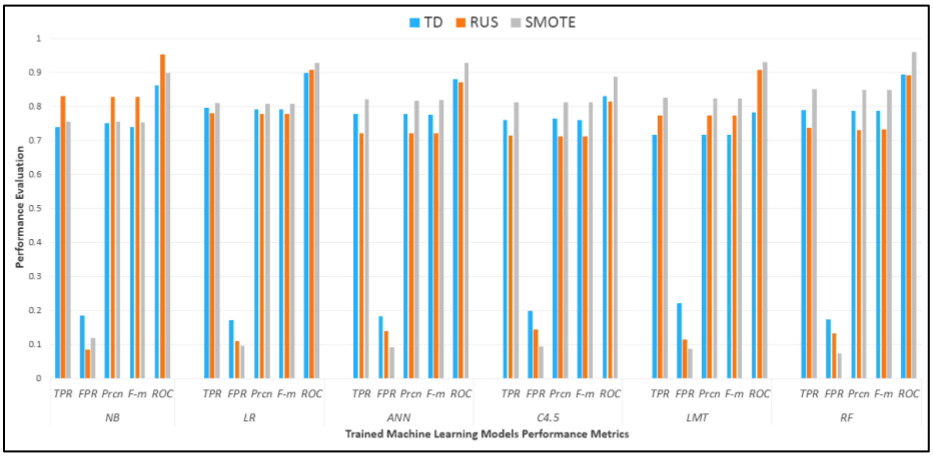


Fig. 3. Comparison of performance metrics across trained models.

3.2 Models Test Results

All models were tested on the same test dataset (N = 2035). When comparing the CCI% for all the models, the CCI% decreased with a maximum degradation of 6.2% when testing the C4.5 model trained on the RUS dataset compared to the same model built on the original TD; this excluded the LMT models which achieved an overall accuracy improvement on test dataset of 6.83% when comparing the SMOTE model to the TD one.

From Fig. 5, it can be observed that in test, RF models achieved the best TPR of 0.770 when trained on TD dataset. LMT model achieved the least TPR of 0.681 when trained on TD dataset. The ROC area across all tested models ranged between 0.786 (for C4.5 on SMOTE) and 0.889 (for LR on TD). These values indicate that also the tested models do not sacrifice a lot of precision to get a good recall on the observed data points.

When observing R, the class of interest, TPR performance results in Fig. 4 show that significant improvements were made in classifying the risk group with the highest level of 0.798 achieved by RF on RUS. RF also achieved the highest TPR improvement in test with a difference of 0.463 between RF on RUS and RF on TD, while NB ranked last, with just 0.121 in minority class TPR improvement between NB on SMOTE and TD. The confusion matrices in Table 3 confirm the above results. The RF model trained on SMOTE correctly classified the highest number of instances (138 of the original 173) in the R group. The model which performed the worst in TPR performance for the minority class R was C4.5 trained on TD, which only correctly classified 43 instances.

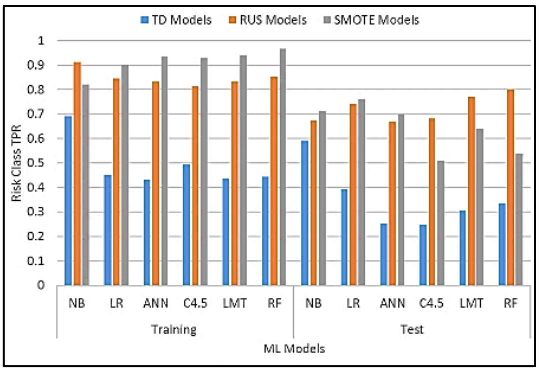
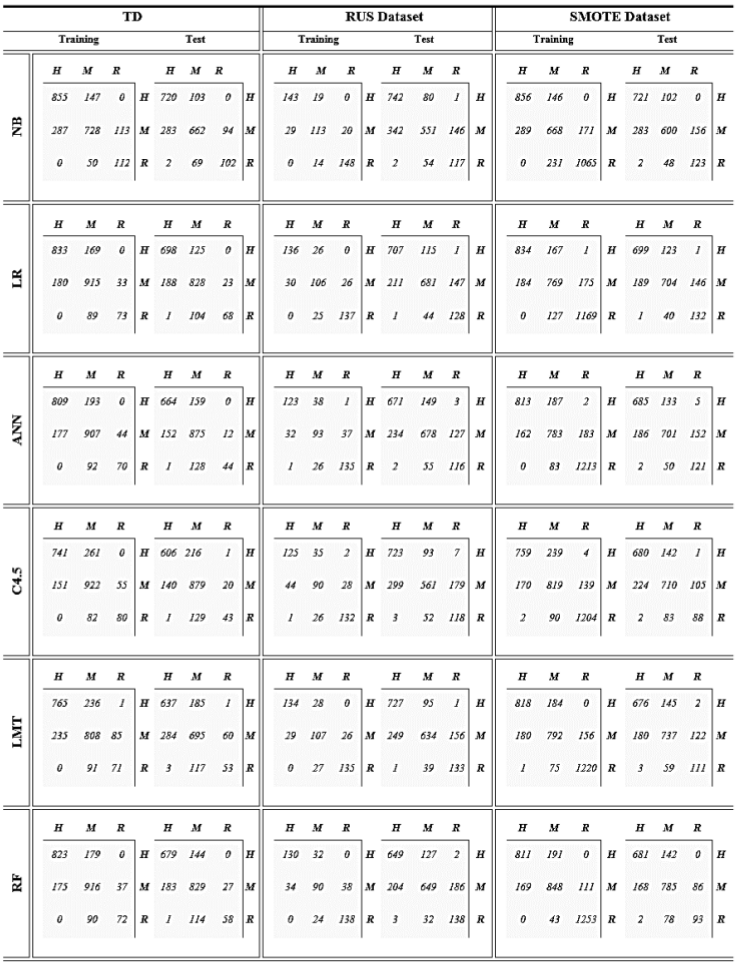


Fig. 4. Risk class TPR performance for trained and tested models per dataset

Table 3. Models Confusion Matrixes Comparison



	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	113	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
	H	M	R		H	M	R		H	M	R		H	M	R		H	M	R		H	M	R	
LR	833	169	0	H	698	125	0	H	136	26	0	H	707	115	1	H	834	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
	H	M	R		H	M	R		H	M	R		H	M	R		H	M	R		H	M	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	135	R	2	55	116	R	0	83	1213	R	2	50	121	R
	H	M	R		H	M	R		H	M	R		H	M	R		H	M	R		H	M	R	
C4.5	741	261	0	H	606	216	1	H	125	35	2	H	723	93	7	H	759	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
	H	M	R		H	M	R		H	M	R		H	M	R		H	M	R		H	M	R	
LMT	765	236	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	156	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
	H	M	R		H	M	R		H	M	R		H	M	R		H	M	R		H	M	R	
RF	823	179	0	H	679	144	0	H	130	32	0	H	649	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	785	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

3.3 Attribute Information Results

When considering the monitored IG for each variable across all datasets (Fig. 6), it is clear that the information gain 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) that correlates directly with the increase in R class 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. 6, SMOTE seems to boost the information within each variable. This, in turn, increases the R class separability from other classes in the training datasets which in turn increases the R class TPR (see Fig. 4). Fig. 7 displays the APA multi-dimensional visualisation which shows the improved R class separability per dataset.

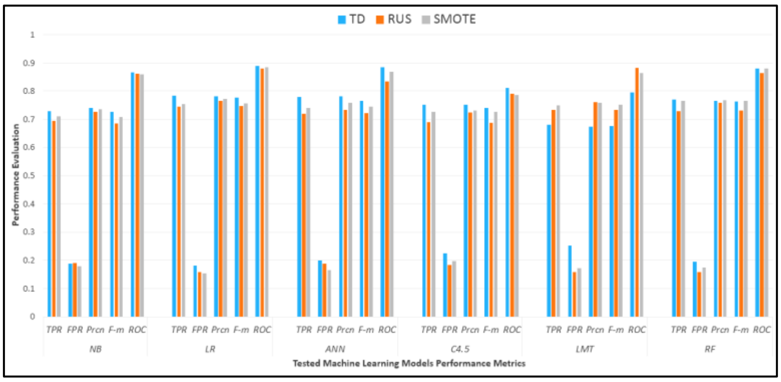


Fig. 5. Comparison of performance metrics across tested models

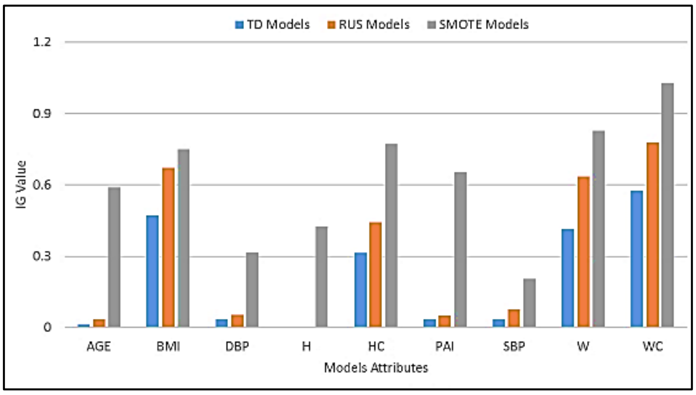


Fig. 6. IG monitoring per variable in each dataset

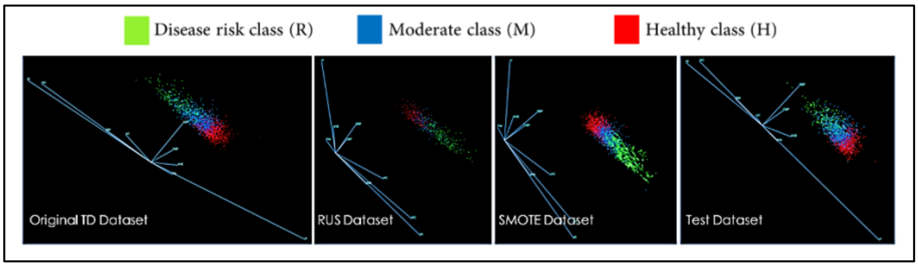
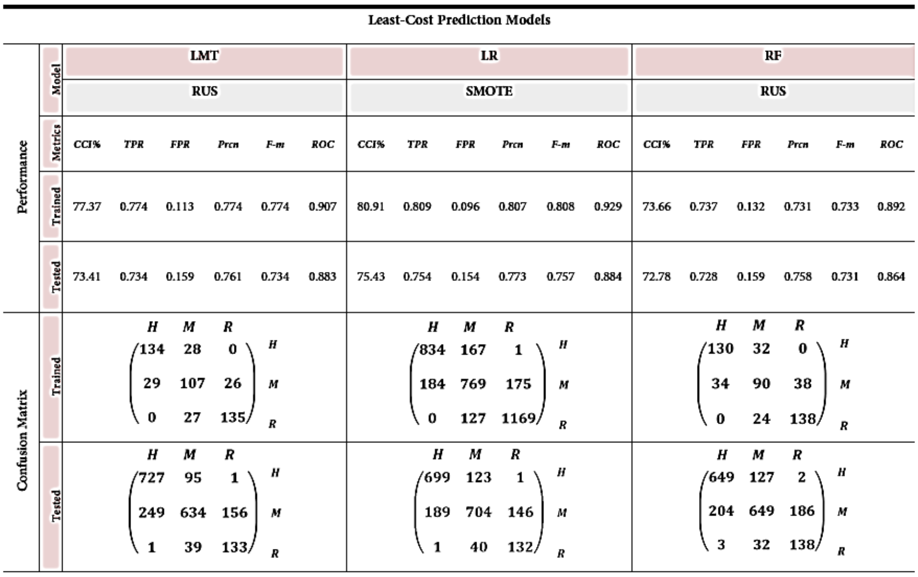


Fig. 7. APA multi-dimensional visualisation

3.4 Domain Experts' Results

During the development of prediction algorithms for use in disease risk-prediction it is important to recognise that the misclassification of subjects, for example false-positive misclassifications, could result in costly and unnecessary follow-up examinations; whereas false-negative misclassifications would result in individuals not receiving interventions to reduce excess VAT. In this particular 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 interventions to improve their lifestyle. Therefore, in common with other scenarios, the best models to adopt would be those which minimise the number of subjects misclassified as at 'risk' in order that they might initiate interventions at an appropriate time. Confusion matrices play an essential role in defining the best-suited model for use in future trials. When analysing the confusion matrices (Table 3), three models were identified as satisfying the domain experts' criteria. These models are reported in Table 4. They did not occupy the highest ranks when their performance metrics were compared to the others.

Table 4. Domain compliant prediction models. N(LMT RUS Trained) = 486; N(LR SMOTE Trained) = 3426; N(RF RUS Trained = 486). N(all Tested) = 2035. For the F-m metric, m=1.



Least-Cost Prediction Models																			
Performance	Model	LMT						LR						RF					
		RUS						SMOTE						RUS					
	Metrics	*CCI%*	*TPR*	*FPR*	*Prcn*	*F-m*	*ROC*	*CCI%*	*TPR*	*FPR*	*Prcn*	*F-m*	*ROC*	*CCI%*	*TPR*	*FPR*	*Prcn*	*F-m*	*ROC*
	Trained	77.37	0.774	0.113	0.774	0.774	0.907	80.91	0.809	0.096	0.807	0.808	0.929	73.66	0.737	0.132	0.731	0.733	0.892
	Tested	73.41	0.734	0.159	0.761	0.734	0.883	75.43	0.754	0.154	0.773	0.757	0.884	72.78	0.728	0.159	0.758	0.731	0.864
Confusion Matrix	Trained	$\begin{array}{cc} \begin{matrix} H & M & R \end{matrix} & \\ \begin{pmatrix} 134 & 28 & 0 \\ 29 & 107 & 26 \\ 0 & 27 & 135 \end{pmatrix} & \begin{matrix} H \\ M \\ R \end{matrix} \end{array}$						$\begin{array}{cc} \begin{matrix} H & M & R \end{matrix} & \\ \begin{pmatrix} 834 & 167 & 1 \\ 184 & 769 & 175 \\ 0 & 127 & 1169 \end{pmatrix} & \begin{matrix} H \\ M \\ R \end{matrix} \end{array}$						$\begin{array}{cc} \begin{matrix} H & M & R \end{matrix} & \\ \begin{pmatrix} 130 & 32 & 0 \\ 34 & 90 & 38 \\ 0 & 24 & 138 \end{pmatrix} & \begin{matrix} H \\ M \\ R \end{matrix} \end{array}$					
	Tested	$\begin{array}{cc} \begin{matrix} H & M & R \end{matrix} & \\ \begin{pmatrix} 727 & 95 & 1 \\ 249 & 634 & 156 \\ 1 & 39 & 133 \end{pmatrix} & \begin{matrix} H \\ M \\ R \end{matrix} \end{array}$						$\begin{array}{cc} \begin{matrix} H & M & R \end{matrix} & \\ \begin{pmatrix} 699 & 123 & 1 \\ 189 & 704 & 146 \\ 1 & 40 & 132 \end{pmatrix} & \begin{matrix} H \\ M \\ R \end{matrix} \end{array}$						$\begin{array}{cc} \begin{matrix} H & M & R \end{matrix} & \\ \begin{pmatrix} 649 & 127 & 2 \\ 204 & 649 & 186 \\ 3 & 32 & 138 \end{pmatrix} & \begin{matrix} H \\ M \\ R \end{matrix} \end{array}$					

4 Discussion

Imbalanced classes have a significant impact on the performance of standard machine learning algorithms. Classification performance in the training phase is severely impacted by class separability. Training the 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, two methods (RUS and SMOTE) have been applied to adjust the class imbalance in the classification training phase at the dataset level. 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], [40]. 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 prediction towards the minority class. The excessive use of such techniques could result in overfitting of the models.

In this study, traditional ML algorithms were sensitive to higher information gains and tended to produce superb performance results in training, but when testing the models, the overall model accuracy often dropped below the training phase performance. The UK Biobank dataset used 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 presented in this study, only three satisfied the domain experts' success criteria for this specific domain problem (LMT and RF built with RUS sampled dataset, and LR built with SMOTE sampled dataset).

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. It would be impractical to compare the results of this study to any other research from the same domain. Nevertheless, this work will be followed by further analyses where additional methods to improve the outcomes will be investigated. Starting from the best-performing methods in this work (LMT and RF), their combination into ensemble learners will first be considered. Future work will also take into account the predictions from this current paper and compare them to the actual incidence of diseases in the same cohort where data is available.

References

1. Yang, Q., Wu, X.: 10 Challenging Problems in Data Mining Research. International Journal of Information Technology & Decision Making 5(4), 597-604 (2006).
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. (eds.) IDEAL 2007, LNCS, vol. 4881, pp 287-295. Springer, Heidelberg (2007).
3. More, A.: Survey of resampling techniques for improving classification performance in unbalanced datasets. arXiv:1608.06048 [stat.AP] (2016).
4. Weiss G.M., 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, A.A.: Imbalanced Data Learning Approaches Review. International Journal of Data Mining & Knowledge Management Process (IJDKP) 3(4), 15-33 (2013).
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, J.M.: A Selective Sampling Method for Imbalanced Data Learning on Support Vector Machines. Ioawa 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 Rao, R., 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, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE : Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 321-357 (2002).
13. Faith, J., Mintram, R., Angelova, M.: Gene expression Targeted projection pursuit for visualizing gene expression data classifications. Bioinformatics 22(21), 2667–2673 (2006).
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, J.R.: Induction of Decision Trees. Machine Learning 1(1), 81-106 (1986).
16. Wang, Y.C., McPherson, K., Marsh, T., Gortmaker, S.L., Brown, M.: Health and economic burden of the projected obesity trends in the USA and the UK. Lancet 378(9793), 815-825 (2011).
17. Sam, S., Mazzone, T.: Adipose tissue changes in obesity and the impact on metabolic function. Translational Research, 164(4), 284-292 (2014).
18. Dattilo, A.M., Kris-Etherton, P.M.: Effects of weight reduction on blood lipids and lipoproteins: a meta-analysis. The American Journal of Clinical Nutrition 56(2), 320–328 (1992).
19. Fox, C.S. *et al.*: Abdominal Visceral and Subcutaneous Adipose Tissue Compartments. Circulation 116(1), 39-48 (2007).
20. Després, J.-P., Lemieux, I., Bergeron, J., Pibarot, P., Mathieu, P., Larose, E., Rodés-Cabau, J., Bertrand, O.F., Poirier, P.: Abdominal Obesity and the Metabolic Syndrome: Contribution to Global Cardiometabolic Risk. Arteriosclerosis, Thrombosis, and Vascular Biology 28(6), 1039-1049 (2008).
21. Chin, S.-H., Kahathuduwa, C.N., Binks, M.: Physical activity and obesity: what we know and what we need to know*. Obesity Reviews 17(12), 1226-1244 (2016).
22. Golabi, P., Bush, H., Younossi, Z.M.: Treatment Strategies for Nonalcoholic Fatty Liver Disease and Nonalcoholic Steatohepatitis. Clinics in Liver Disease 21(4), 739-753 (2017).
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 52(10), 2532-2538 (2003).
24. Brouwers, B., Hesselink, M.K.C., Schrauwen, P., Schrauwen-Hinderling, V.B.: Effects of exercise training on intrahepatic lipid content in humans. Diabetologia 59(10), 2068-2079 (2016).
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 12(3), e1001779 (2015).
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 5(6), 691-705 (2011).
28. Drummond, C., Holte, R.C.: 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., Schutze, H.: Introduction to Information Retrieval. Natural Language Engineering 16(1), 100–103 (2010).
30. Zhang H.: The Optimality of Naive Bayes. American Association for Artificial Intelligence (2004).
31. Landwehr N., Hall, M., Frank, E.: Logistic Model Trees. Machine Learning 59(1-2), 161–205, 2005.
32. Ayer, T., Chhatwal, F., Alagoz, O., Kahn, C.E., Woods, R.W., Burnside, E.S.: Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation. Radio Graphics 30(1), 13–22 (2010).
33. Quinlan J.R.: Improved Use of Continuous Attributes in C4.5. Journal of Artificial Intelligence Research 4, 77-90 (1996).
34. Witten, I.H., Frank, E.: Data Mining, Practical Machine Learning Tools and Techniques. 2nd edn. Elsevier Inc (2005).
35. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, 2000.

36. Jonsdottir, T., Hvannberg, E.T., Sigurdsson, H., Sigurdsson, S.: The feasibility of constructing a Predictive Outcome Model for breast cancer using the tools of data mining. Expert Systems with Applications 34(1), 108–118 (2008).
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 2(7), 1-5 (2011).
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, J.R. *et al.*: Visceral adipose tissue, thigh adiposity and liver fat fraction: a cross sectional analysis of the UK Biobank. UK Biobank (2019).
40. 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.