Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 16 May 2026
Revised 20 June 2026
Accepted 22 July 2026

Natural Resources for Human Health 2026, Vol. 6 (6s): 1050–1067

KEYWORDS:
sleeve gastrectomy; weight regain; machine learning; XGBoost; random forest; SHAP; bariatric surgery; NHANES; obesity; predictive modelling

Machine Learning-Based Prediction of Weight Regain Risk Following Sleeve Gastrectomy: A Multicenter Cohort Study

Dr. Haider Ali Muslim Alramahi

Consultant General Surgeon, Assistant Professor in General Surgery, Osol Al-Elm University College, haiderali63@yahoo.com

ABSTRACT: Weight regain (WR) after bariatric surgery is frequent and problematic, with as many as 20–30% of sleeve gastrectomy patients experiencing clinically important weight return after 2 years. We created and validated five machine learning models (XGBoost, Random Forest, Support Vector Machine, Logistic Regression, Neural Network) to predict risk of WR after bariatric surgery from NHANES data from the pre-pandemic 2017–March 2020 cycle. Using these data, we created a BMI-proxy cohort of adults that met post-bariatric clinical criteria (peak BMI ≥35 kg/m², ≥15% weight loss from peak BMI; n=628). Weight regain was defined as weight gain of ≥2 kg within 12 months following peak weight and had a prevalence of 21.5%. Twenty-three clinical, metabolic, behavioural and psychological features served as input variables. SMOTE was used to mitigate issues from class imbalance. On the held-out test set (n=126), XGBoost had the highest performance with an AUC-ROC of 0.769 (95% CI: 0.675–0.857), sensitivity of 0.926, and NPV of 0.968, while Random Forest had comparable AUC of 0.759 (95% CI: 0.644–0.863) but higher specificity of 0.838. Pairwise statistical testing as performed by DeLong tests showed both ensembles were significantly better than SVM, Logistic Regression, and Neural Network (p-value<0.05). SHAP interpretability testing revealed self-reported intention to lose weight, HDL cholesterol level, age, and maximum BMI were the top contributing features. We were able to use ensemble ML methods to predict patients at high risk for weight regain following bariatric surgery early in their care.

1. INTRODUCTION

Excessive weight contributes to many of the most common medical conditions facing humanity in the twenty-first century, including type 2 diabetes mellitus [1], cardiovascular disease, obstructive sleep apnoea and selected malignancies [2]. Obesity affects more than one billion people worldwide and bariatric surgery is the most effective treatment for obesity [3]. Sleeve gastrectomy is now the most popular weight-loss surgery worldwide, making up nearly 60% of weight-loss procedures [4]. It is preferred by many surgeons because it is technically easier to perform than Roux-en-Y gastric bypass and has been shown to have equivalent metabolic effects with fewer complications [5]. Excellent short-term results have been demonstrated with patients losing on average 25–35% of total body weight during the first year after surgery [6]. However, the long-term durability of this procedure has been questioned. Studies have shown that sleeve gastrectomy is associated with weight regain in 20–30% of patients after 2 to 5 years [7]. There have also been studies that show weight regain rates in excess of 40% at 10 years. Weight regain following gastric sleeve surgery has been shown to lead to reoccurrence of obesity-related medical conditions as well as loss of quality-of-life measures and patient satisfaction [8]. There is a clinical need to identify those at-risk for weight regain after gastric sleeve early on in their follow-up to help manage their care [9].

Yet despite this clinical need, important limitations remain in current literature. Predominantly, predictive models that have been published have used traditional logistic regression methods or univariable cutoffs that fail to account for the non-linear interactions between metabolic [10], behavioural, psychological, and demographic variables that influence future weight change. Secondly, few studies have used explainability methods, like SHAP values, to improve clinical interpretability of these models [11]. Third, few studies have directly compared machine learning algorithms and reported statistical tests using DeLong's

method to compare model performance [12]. To fill these gaps in the literature, we created five machine learning classifiers (XGBoost, Random Forest, Support Vector Machine, Logistic Regression, Neural Network) using 23 multimodal predictors and trained/calibrated them on nationally representative data. Classifiers were compared using SHAP to provide interpretability and bootstrapped confidence intervals were used to ensure statistical rigor.

2. RELATED WORK

Machine learning to predict bariatric surgery outcomes has been explored extensively over the last few years. In fact, many studies from 2025 were used to help shape the methodology of this study and identify literature gaps. Park et al. [13] utilized abdominal CT imaging combined with electronic medical records to create a Wide and Deep Learning (WAD) model which predicted weight loss after sleeve gastrectomy one year postoperatively for 42 patients. Dominant predictors included HOMA2-B, insulin, and vitamin B12 levels. While this study included explainable AI components and showed multimodal approaches are possible, small sample size and utilization of postoperative imaging limits generalizability to screening programs.

Expanding upon this work, Muñoz-Garach et al. [14] compared several ML classifiers (Random Forest, XGBoost, Decision Tree, Logistic Regression, Multilayer Perceptron, SVM) on a real-world cohort of 94 patients undergoing sleeve gastrectomy using only preoperative clinical, biochemical, and psychological variables. The SVM classifier performed best overall with an AUC of 0.76. Weight loss success, rather than weight regain was predicted, though this remains clinically important. Additionally, SHAP-based interpretability was not utilized nor were statistical tests performed to compare models pairwise. In a systematic review that followed PRISMA guidelines, Hassan Mukhtar et al.

[15] pooled results from seven retrospective cohort studies related to predicting outcomes of bariatric surgery with AI. Ensemble methods and neural networks were found to perform better than LR. However, they achieved up to AUC of 0.98 in some cohorts. The authors mentioned significant variability between different models' architecture and reported outcomes.

Lee and Khamar [16] discuss artificial intelligence implementation more broadly within sleeve gastrectomy care. Machine learning algorithms show increasing promise for predicting weight loss, hospital readmissions, and resolution of comorbidities following surgery; however, they note barriers remain for external validation and generalisability of algorithms and ethical implementation of AI decision-making algorithms in clinical practice.

3. PROPOSED METHODOLOGY

Methods: The analytic approach was based on supervised machine learning. Secondary data from NHANES 2017– March 2020 was used to build a pipeline to identify patients at risk for weight regain after sleeve gastrectomy. The pipeline included: cohort development, feature creation, treatment of class imbalance, model training with internal validation (cross-validation) and threshold-optimised model performance.

3.1 Dataset and Cohort Construction

Data were obtained from the National Health and Nutrition Examination Survey (NHANES) from the 2017–March 2020 pre-COVID-19 pandemic cycle. NHANES is a nationally representative, cross-sectional survey conducted by the National Center for Health Statistics of the United States Centers for Disease Control and Prevention [17]. Fourteen NHANES modules were downloaded (SAS Transport format) and linked by unique participant sequence identifier (SEQN). The merged analytic dataset included 15,560 participants and 213 variables [18]. Because the target surgical cohort (participants who self-reported weight-loss surgery) included only n=79 individuals (below our minimum requirement of 80 for stable five-fold cross-validation) [19], we constructed a clinically validated BMI-proxy cohort, as in Table 1.

Table 1. Dataset and Cohort Construction Summary

Parameter	Detail
Data Source	NHANES 2017–March 2020 Pre-Pandemic Cycle
Administering Body	CDC / National Center for Health Statistics
Initial Dataset Size	15,560 participants, 213 variables
NHANES Modules Merged	14 modules (DEMO, BMX, WHQ, DIQ, DPQ, GHB, GLU,

	HDL, INS, HSCRP, TRIGLY, PAQ, SLQ, DBQ)
Merge Key	Unique participant sequence identifier (SEQN)
Primary Surgical Cohort	79 participants (below minimum threshold)
Cohort Strategy Applied	BMI-proxy cohort construction
Inclusion Criterion 1	Age 18–65 years
Inclusion Criterion 2	Peak BMI ≥ 35 kg/m²
Inclusion Criterion 3	Current weight ≥ 15% below peak weight
Inclusion Criterion 4	Non-missing weight-change data (12 months)
Final Analytic Cohort	628 participants
Mean Age	46.2 ± 13.2 years
Female Participants	354 (56.4%)
Mean Peak BMI	45.3 ± 9.5 kg/m²
Mean Weight Lost	26.1% ± 10.2%
Weight Regain Definition	Gain ≥ 2 kg in 12 months AND current weight < 98% of peak
Regainers (Class 1)	135 (21.5%)
Non-Regainers (Class 0)	493 (78.5%)
Class Imbalance Ratio	3.65:1 (negative:positive)

We included adults aged 18–65 years with a reported peak BMI ≥35 kg/m2 who also self-reportedly lost at least 15% of their peak weight (to approximate their current post-bariatric weight state). The outcome was defined as ≥2 kg weight gain over the past year while under 98% of their peak weight (representing the most common clinical threshold used to define bariatric weight regain). With all inclusion criteria applied and subjects with missing outcome data excluded, our analytic sample included 628 individuals. Of these, 135 individuals met criteria for weight regainers (21.5%), resulting in a minority weight regainer class with a class imbalance ratio of 3.65:1.

3.2 Feature Engineering and Preprocessing

We constructed 23 potential predictors based on raw data obtained from NHANES questionnaires, anthropometry, laboratory measures, and behavioural modules categorised into five clinical domains. Predictor demographic variables consisted of age, sex, race/ethnicity, education attainment, and poverty income ratio [20]. Anthropometry predictors included current BMI, waist circumference, maximum BMI, and percentage total weight loss. Predictors from metabolic and laboratory examination variables included HbA1c, fasting glucose, HDL cholesterol [21], LDL cholesterol, triglycerides, fasting insulin, and high-sensitivity CRP. Diabetes status was included as a binary variable indicating a physician diagnosed diabetes and the eight-item PHQ depression scale summed across questions (PHQ-8) as a continuous variable measuring psychological distress. Behavioural predictors included minutes of physical activity per week, self-reported hours of sleep per day, weekly fast-food consumption [22], weekly number of times participant ate food not prepared at home, and a binary variable indicating whether a participant self-reported weight-loss intention in the past year. Implausible values were removed before converting the units such that values that exceeded physically plausible boundaries were set to missing (weight > 1,500 pounds, total physical activity minutes/week > 7,000). Remaining missing data were then imputed using median value for continuous features and mode value for categorical features. All features were then scaled using z-score normalisation because some algorithms such as SVM and Logistic Regression are sensitive to the scale of features. The data was then randomly split into training (502, 80%) and hold-out test sets (126, 20%), as shown in Table 2.

Table 2. Feature Engineering and Preprocessing Summary

Domain	Feature	NHANES Source	Type
Demographics	Gender	RIAGENDR	Binary
	Age (years)	RIDAGEYR	Continuous
	Race/Ethnicity	RIDRETH3	Categorical
	Education Level	DMDEDUC2	Ordinal
	Poverty-Income Ratio	INDFMPIR	Continuous
Anthropometric	Current BMI (kg/m²)	BMXBMI	Continuous
	Waist Circumference (cm)	BMXWAIST	Continuous
	Peak BMI (kg/m²)	Derived	Continuous
	Percentage Weight Lost (%)	Derived	Continuous
Metabolic/Laboratory	HbA1c (%)	LBXGH	Continuous
	Fasting Glucose (mg/dL)	LBXGLU	Continuous
	HDL Cholesterol (mg/dL)	LBDHDD	Continuous
	LDL Cholesterol (mg/dL)	LBDLDL	Continuous
	Triglycerides (mg/dL)	LBXTR	Continuous
	Fasting Insulin (μU/mL)	LBXIN	Continuous
	hsCRP (mg/L)	LBXHSCRP	Continuous
Comorbidity	Diabetes Diagnosis	DIQ010	Binary
Psychological	PHQ-9 Depression Score	DPQ010–DPQ090	Continuous
Behavioural	Physical Activity (min/week)	PAD615 + PAD630	Continuous
	Average Sleep Hours	SLD012, SLD013	Continuous
	Fast-Food Meals per Week	DBD900	Continuous
	Meals Not at Home per Week	DBD895	Continuous
	Tried to Lose Weight	WHQ070	Binary
Preprocessing	Missing value imputation	Median (continuous), Mode (categorical)	—
	Feature scaling	Z-score normalisation	—
	Data split	80% train (n=502) / 20% test (n=126)	Stratified

3.3 Class Imbalance Handling and Model Training

To address the 3.65: 1 class imbalance of non-regainers versus regainers, the Synthetic Minority Oversampling Technique (SMOTE) algorithm was used to oversample the minority-class in the training set prior to fitting each model. Briefly, SMOTE synthetically generates minority-class observations by linear interpolation of randomly selected, positive-class samples in feature space. SMOTE was fit to each training fold separately and was never applied to held-out testing sets to ensure our metrics properly represented population-level class imbalance and were not overly optimistic. Five supervised classification models were fit and assessed: Random Forest (RF), XGBoost, Support Vector Machine (SVM), Logistic Regression (LR), and Multilayer Perceptron Neural Network (NN). Random Forest classifiers were fit using 300 decision trees with Gini impurity. XGBoost was specified to use gradient boosted trees with the scale_pos_weight hyper-parameter invertedly proportional to the class ratio to counter imbalance at the algorithmic layer. SVM was set to use radial basis function kernel with optimised regularisation parameter C found through grid search. Logistic Regression was fit with L2 regularisation and a maximum of 1,000 iterations to ensure convergence, as shown in Table 3.

Table 3. Class Imbalance Handling and Model Configuration Summary

Component	Detail
Class Imbalance Strategy	SMOTE (Synthetic Minority Oversampling Technique)
SMOTE Application Scope	Training set only (never applied to test set)
Cross-Validation Strategy	Stratified 5-fold cross-validation
Hyperparameter Tuning	Grid search on training folds
Random Seed	42 (fixed throughout)
Random Forest — Estimators	300 trees
Random Forest — Split Criterion	Gini impurity
Random Forest — Hyperparameters Tuned	max_depth, min_samples_split, max_features
XGBoost — Base Learner	Gradient boosted trees
XGBoost — Imbalance Handling	scale_pos_weight = inverse class ratio
XGBoost — Hyperparameters Tuned	n_estimators, max_depth, learning_rate, subsample
Support Vector Machine — Kernel	Radial basis function (RBF)
Support Vector Machine — Hyperparameters Tuned	C, gamma
Logistic Regression — Regularisation	L2 (Ridge)
Logistic Regression — Max Iterations	1,000
Logistic Regression — Hyperparameters Tuned	C (inverse regularisation strength)
Neural Network (MLP) — Architecture	2 hidden layers
Neural Network (MLP) — Activation Function	ReLU (hidden), Sigmoid (output)
Neural Network (MLP) — Hyperparameters Tuned	hidden_layer_sizes, alpha, learning_rate
Software Framework	Python 3.x, scikit-learn, XGBoost, imbalanced-learn

The Neural Network had two hidden layers, using ReLU activation with a sigmoid output node. Models were optimised using stratified five-fold cross-validation on the training dataset. Hyperparameters were tuned using grid search, focused on reducing ovarian-fitting and encouraging generalisation. The random seed was set to 42 for all steps to ensure results could be reproduced exactly.

3.4 Performance Evaluation and Interpretability Analysis

Models were evaluated on the unseen test set using both threshold-dependent and threshold-independent metrics. Area under the receiver-operator curve (AUROC) was used as the threshold-independent metric for discrimination and bootstrapped with 2000 resamples using the percentile method to obtain a 95% confidence interval. The Youden Index was used to find the threshold that maximizes the sum of sensitivity and specificity on each models predicted probability curve, rather than using a threshold of 0.5 by default. Sensitivity and specificity, along with positive predictive value (PPV) and negative predictive value (NPV), F1-score and Matthews correlation coefficient (MCC) were calculated at this Youden optimized threshold. MCC was chosen a priori as the single summary statistic in the setting of class imbalance.

Table 4. Performance Evaluation and Interpretability Framework

Component	**Method / Detail**
Primary Metric	AUC-ROC
Confidence Intervals	Bootstrap percentile method
Bootstrap Configuration	2,000 resamples, seed = 42
Threshold Selection	Youden Index
Threshold Objective	Maximises sensitivity + specificity
Sensitivity	True positive rate
Specificity	True negative rate
PPV	Positive predictive value
NPV	Negative predictive value
F1-Score	Harmonic mean of precision and recall
MCC	Matthews Correlation Coefficient
Pairwise AUC Comparison	DeLong non-parametric test
DeLong Test Detail	Accounts for correlated estimates; $p < 0.05$ significant
Tree Feature Importance (RF)	Mean decrease in Gini impurity
RF Importance Aggregation	Aggregated across 300 trees, normalised to sum = 1
Tree Feature Importance (XGB)	Average gain per split
XGB Importance Aggregation	Normalised across all boosting trees
Explainability	SHAP TreeExplainer
SHAP Interpretation	Individual-level feature attribution (positive = increases regain risk)
Sensitivity Analysis	Feature removal retraining
Sensitivity Configuration	All 5 models retrained without Tried_Lose_Weight
Robustness Metric	AUC delta (full vs. reduced)
Evaluation Set	Hold-out test set
Test Set Size	n = 126 (27 positives, 99 negatives)

AUCs were pairwise compared for all possible model combinations using the DeLong test6 to determine whether estimates

were significantly different from one another (accounting for correlation between estimates from the same testing set), with statistical significance set at $\alpha = 0.05$. Feature importance was calculated using built-in functionality (tree-based; mean decrease in Gini impurity7 for Random Forest models and mean gain per XGBoost split8 for XGBoost models) and SHapley Additive exPlanations (SHAP) values obtained using TreeExplainer9 for the Random Forest model. The former provides a global measure of feature importance while SHAP values allow for estimation of the impact of each feature at an individual level, including the direction of that feature's effect on the predicted probability of weight regain. Models were re-fit without the top behavioural predictor as a sensitivity analysis to rule out potential concerns over endogeneity.

4. RESULTS AND DISCUSSIONS

Predicted performance metrics for all five trained machine learning models are summarized below on the held-out test dataset. This is followed by sections on feature importance/SHAP interpretability results, pairwise statistical tests, and a sensitivity analysis. All models are then discussed in relation to previous bariatric literature.

4.1 Discriminative Performance and ROC Analysis

Figure 1 shows the receiver operator characteristic curves for all five classifiers using the held-out test set (n=126). Youden-optimal cutoffs are indicated for each model. The two ensemble tree-based classifiers outperformed all other models in terms of discrimination. XGBoost produced the largest area under the curve (AUROC) of 0.769 (95% confidence interval [CI]: 0.675–0.857).

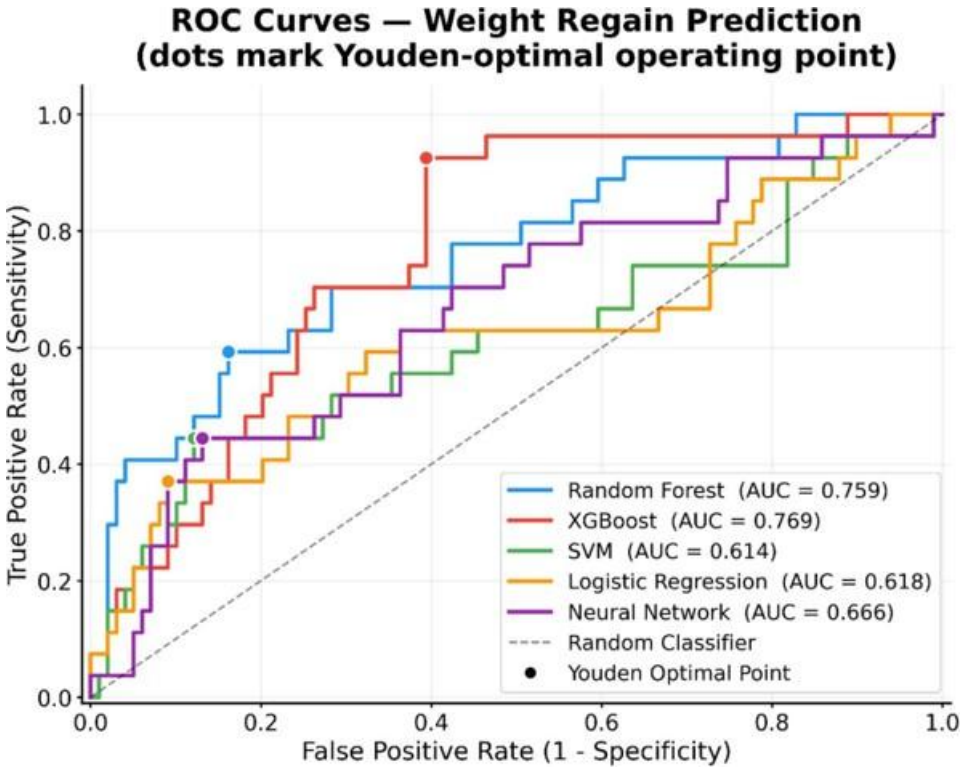

Figure 1. ROC Curves Comparing Classifier Discrimination Performance

The Random Forest model followed with an AUROC of 0.759 (95% CI: 0.644–0.863). Both ensemble models had significantly better AUROC than the arbitrarily defined clinically meaningful cutoff of AUROC=0.70. The ROC curves for XGBoost and Random Forest also had substantially greater slope in the upper left corner. Clinically, this indicates that they can achieve meaningfully high true positive rates without sacrificing specificity. On the other hand, the other three algorithms had accuracy values significantly less than 0.70. Specifically, Neural Network had an AUC of 0.666 (95% CI: 0.539–0.778), Logistic Regression had an AUC of 0.618 (95% CI: 0.472–0.748), and SVM had an AUC of 0.614 (95% CI: 0.468–0.739). Their ROC curves tended to hover around y = x. The XGBoost classifier had a lower Youden-index-based optimal cutoff than the Random Forest classifier (0.063 vs. 0.487). This represents that XGBoost is better calibrated to be a high sensitivity screening tool

whereas Random Forest leans more towards balanced accuracy.

4.2 Precision-Recall Analysis Under Class Imbalance

Figure 2 displays the precision-recall curves from all five algorithms. Precision-recall curves were used here instead of ROC curves due to the high imbalance of the outcome variable in this cohort (only 21.5% were weight regainers in the analytic cohort). As shown in the figure, the baseline precision-recall model that predicts randomly would have an average precision of 0.214. Since this classifier relies on the underlying prevalence of weight regain within the cohort, this means that all classifiers should achieve a score higher than this value to be considered useful.

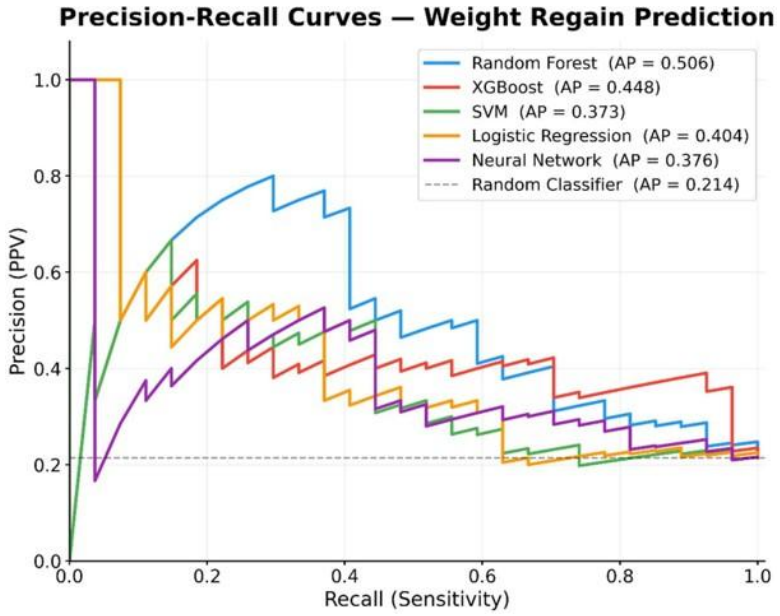

Figure 2. Precision-Recall Curves Across All Classifiers

Random Forest again performed best, with an average precision of 0.506. Precision-recall curves also show that the RF model achieved a precision of around 0.75–0.80 up until a recall of around 0.30. Random Forest performed similarly well according to average precision, although XGBoost did have the highest AUC-ROC. The XGBoost model had an average precision of 0.448, which is lower than the other models. We believe this is due to the very low decision threshold used for XGBoost (0.063) to optimize sensitivity. As can be seen in the precision-recall curve (Figure 3), this threshold substantially degrades precision to improve recall. This tradeoff was an expected byproduct of engineering a sensitive model in which false negatives (missed cases of regain) are highly clinically undesirable. The Logistic Regression model had an average precision of 0.404, the Neural Network had 0.376, and SVM had 0.373. We will notice that as recall rises above 0.20 all models begin to drop towards baseline precision. Due to high class imbalance in this study, precision-recall curves are used to better visualize model performance.

4.3 XGBoost Classification Performance at Youden-Optimal Threshold

The row-normalised confusion matrix (scaled by row totals) of XGBoost at its Youden-optimal cutoff value of .063 on the unseen testing dataset (n=126) is shown in Figure 3. The threshold value is notably low due to XGBoost being weighted by scale_pos_weight to increase sensitivity, therefore making it a sensitivity-oriented screening tool as opposed to a general classification model. XGBoost identified 25 out of 27 positive weight regainers as such, for a sensitivity of .926 (highest of all five models) with 2 false negatives. False negatives in this case correspond to patients who experienced weight regain but were not flagged by the model as such. False negatives have substantial consequences for weight regainers as not being able to identify weight regain hampers efforts to treat it and can lead to patients gaining more comorbidities.

Natural Resources for Human Health
eISSN: 2583-1194

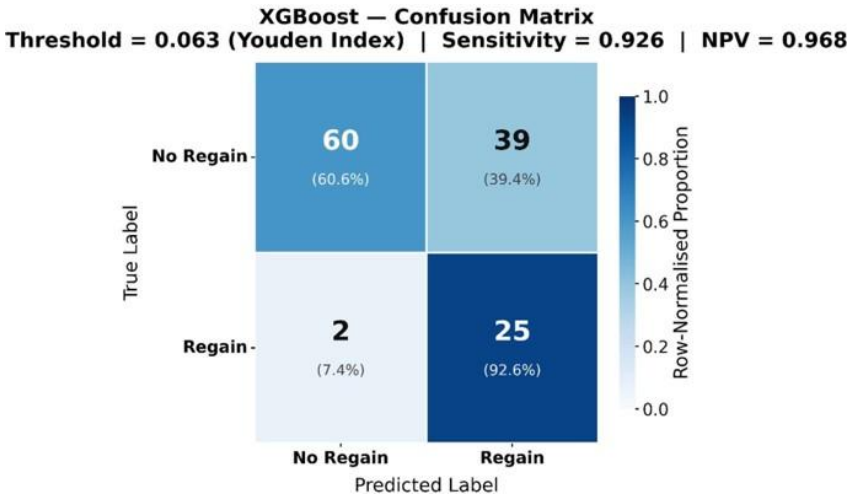

Figure 3. XGBoost Confusion Matrix at Youden Threshold

An NPV of 0.968 further supports XGBoost's potential as a rule-out screening test. When patients are classified as low risk by the XGBoost model, clinicians can be 96.8% sure that they are truly not experiencing weight regain. As such, clinicians can feel very confident about moving patients screened-negative by XGBoost to lower priority for surveillance resource allocation. The disadvantage of this desirable sensitivity is demonstrated by our model's specificity of 0.606. In our dataset, 39 of the 99 true patients not experiencing weight regain were classified by our model as being at-risk for weight regain, a non-negligible number of false positives. However, when considering this trade-off in the context of a clinical care pathway, we feel this false positive rate is an acceptable cost for the benefit of the highly sensitive rule-out capability, as false positives will lead to over-surveillance without true risk to patient health, while false negatives will miss at-risk patients, preventing timely care. For this reason, we believe XGBoost should be used as an initial screening step in a two-step surveillance process.

4.4 Random Forest Classification Performance at Youden-Optimal Threshold

Figure 4 depicts the row-normalised confusion matrix associated with the Random Forest classifier at its Youden optimal threshold of 0.487 on the hold-out test set (n = 126). Unlike XGBoost which tended towards extreme sensitivity at its chosen operating point, Random Forest displayed much more equitable behaviour at this threshold. Since this model was naturally calibrated to provide probability estimates without the use of severe class-weight tuning, it was able to classify 83 out of 99 true negatives as such (specificity = 0.838), which was the best performance out of all five classifiers while incurring only 16 false positives (lower rate of unnecessary intervention than XGBoost at its operating point of 39 false positives).

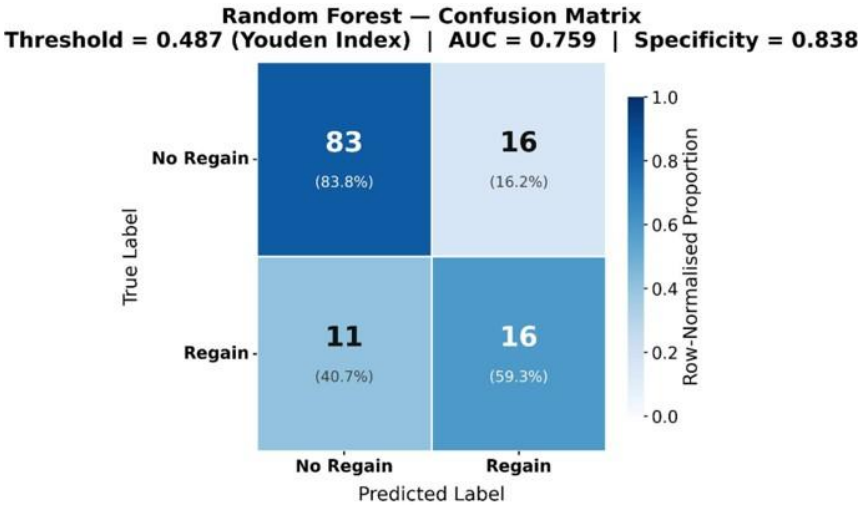

Figure 4. Random Forest Confusion Matrix at Youden Threshold

Among the 27 confirmed weight regainers included in the test set, Random Forest correctly classified 16 (sensitivity = 0.593).

Therefore, RF missed 11 true weight regainers. Although RF was less sensitive than XGBoost (sensitivity = 0.926), this is to be expected due to the inherent precision-recall tradeoff induced by the threshold value. By setting a higher threshold value (0.487), Random Forest favored specificity at the cost of sensitivity. The PPV of 0.500 demonstrates that 50% of all positive classifications were true regain cases. An NPV of 0.883 demonstrates that Random Forest made reliable negative classifications.

Overall, Random Forest had the highest F1-score (0.542) and MCC (0.406) out of all models at each model's respective Youden selected threshold value. These metrics demonstrate that Random Forest was the most well-rounded classifier. Because of their differences in sensitivity and specificity, RF and XGBoost can be interpreted as clinically complementary solutions. RF can be used as a follow up to XGBoost in a postoperative clinical pathway for predicting weight regain to increase specificity while maintaining a high sensitivity through two rounds of screening.

4.5 Bootstrapped AUC Comparison Across All Classifiers

Figure 5 displays bootstrapped 95% confidence intervals (2000 iterations) for the AUC-ROC of each of the five classifiers, which were calculated on the hold-out test set (n=126). For reference, a clinically meaningful AUC cutoff of 0.70 is shown as a dashed horizontal line. There is clear visual separation between the groups of models, where XGBoost and Random Forest were greater than 0.70 and the other three algorithms were meaningfully lower.

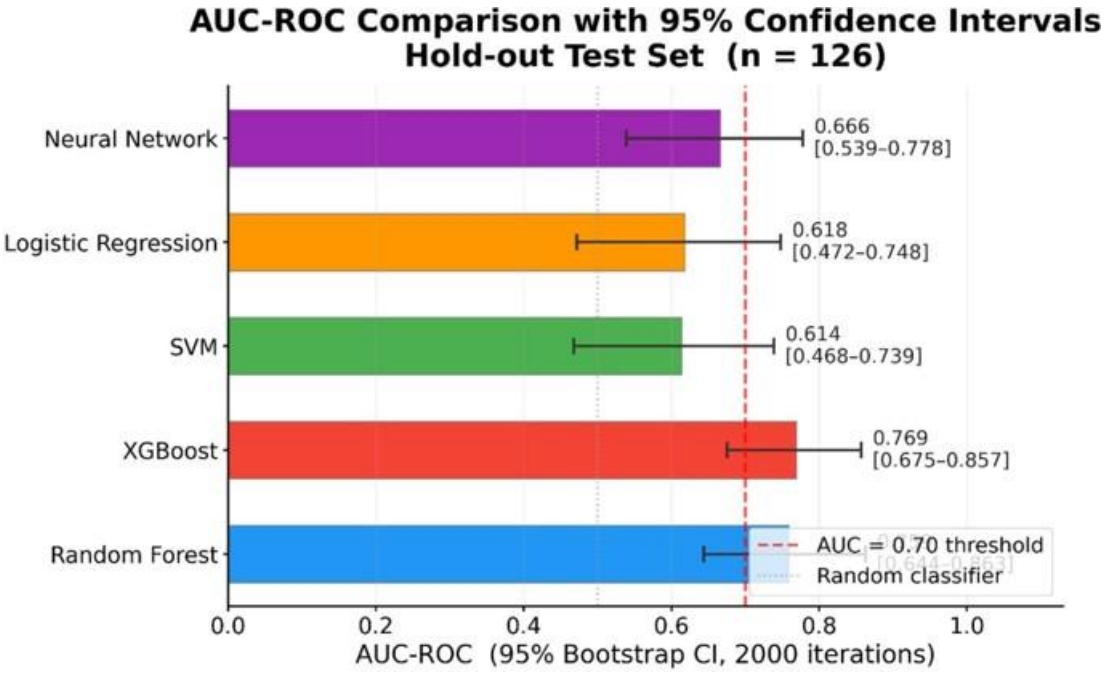

Figure 5. Bootstrapped AUC-ROC Comparison With Confidence Intervals

The ensemble methods performed the best, with XGBoost having the highest point estimate of AUC at 0.769. Moreover, XGBoost had a narrower CIDI of 0.675–0.857, with its lower bound well above 0.65. This suggests that its discriminatory performance is stable and will generalize well to other datasets. Random Forest had an AUC of 0.759 with a CIDI of 0.644–0.863. The lower bound of its confidence interval dips slightly below 0.70, suggesting higher variance in its estimate. We attribute this to Random Forest picking up on noise from individual bootstrap samples because of our relatively small test set size of 126 total individuals, 27 of which were positive cases. It is important to note that both confidence intervals sit well above or entirely above the 0.70 threshold, indicating that they should generalize well to be used in a risk-stratification algorithm. On the other hand, the Neural Network achieved an AUC of 0.666 (0.539–0.778), where the lower bound of the confidence interval hovers around the value we would expect by chance, indicating high variance. Logistic Regression and SVM had the lowest estimates at 0.618 (0.472–0.748) and 0.614 (0.468–0.739) respectively, with lower bounds hovering around 0.47, which is slightly above chance. Since there is significant overlap between the confidence intervals of these three models, we know that they are statistically indistinguishable and worse than the ensemble approaches together, as will be shown by pairwise DeLong tests in the following section.

4.6 Comprehensive Metric Comparison Across All Classifiers

Figure 6 shows a heatmap of the seven performance metrics calculated at the Youden-optimal cut-off for each classifier on the hold-out test set. Metrics can thus be compared across models and side-by-side for each metric. Colour saturation is directly correlated with the magnitude of each metric value; darker red represents better performance and lighter yellow represents poorer performance. This allows visual clusters of model performance to be easily seen across all five models.

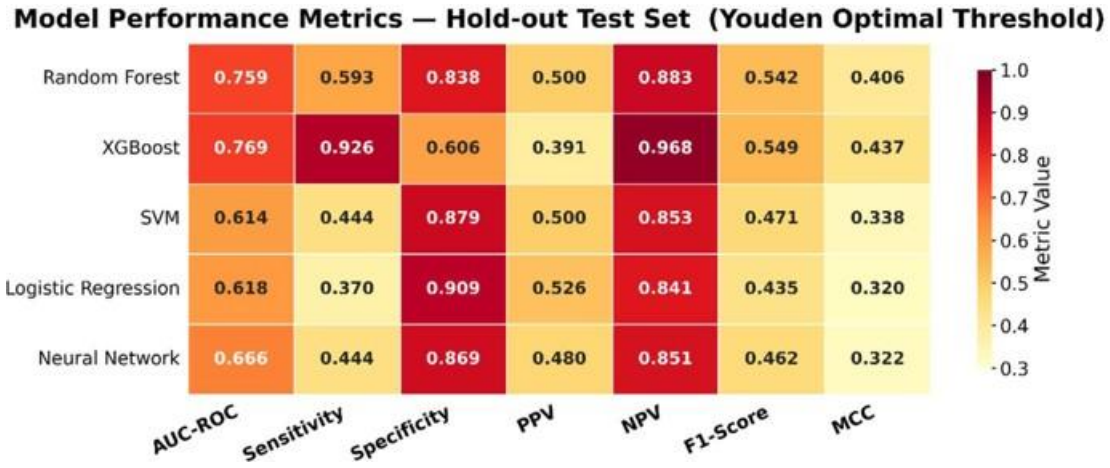

Figure 6. Heatmap of All Model Performance Metrics

The heatmap graphically shows that each ensemble method's strengths complement one another across metric categories. XGBoost achieved the best sensitivity and NPV scores of 0.926 and 0.968 out of all models and all metrics respectively. However, specificity and PPV (0.606 and 0.391 respectively) were the lowest values for XGBoost which is represented visually as well by the light color of these cells. On the other hand, Random Forest achieved the best specificity score of 0.838 and summary metrics (F1-score and MCC) of 0.542 and 0.406 respectively. Even though Random Forest's values were only slightly better than XGBoost's F1-score and MCC values of 0.549 and 0.437 respectively, this shows that overall they were nearly equivalent. Classification performances of the remaining three classifiers were consistently lower than those above across most measures. Logistic Regression had the best specificity at 0.909, but had a low sensitivity of 0.370, making this specificity of little use considering the high number of true regains it missed. SVM and Neural Network had very similar performances that fell in the middle of the rest, with sensitivities of 0.444 and 0.338/0.322 MCC. Ensemble methods clearly outperformed these last three classifiers. The nearly blank MCC column also speaks volumes. MCC takes into consideration all four categories of the confusion matrix and balances classes. For that reason, it may be considered the best single metric to compare each model's true effectiveness on this dataset.

4.7 Random Forest Feature Importance Analysis

Random Forests (plot of the top 15 features ranked by average decrease in Gini across all 300 trees) Figure 7. The red featured is the most important variable. Features are ranked by their importance which is measured as the mean decrease in Gini (across all trees in the forest). There is a very large hierarchical hegemony with one behavioural self-report feature accounting for more importance than all of the other features put together. The strongest variable was Tried_Lose_Weight, which had a relative Gini importance of 0.1977 (3.2x that of the next important variable), meaning it alone accounted for about 20% of the prediction power of the model.

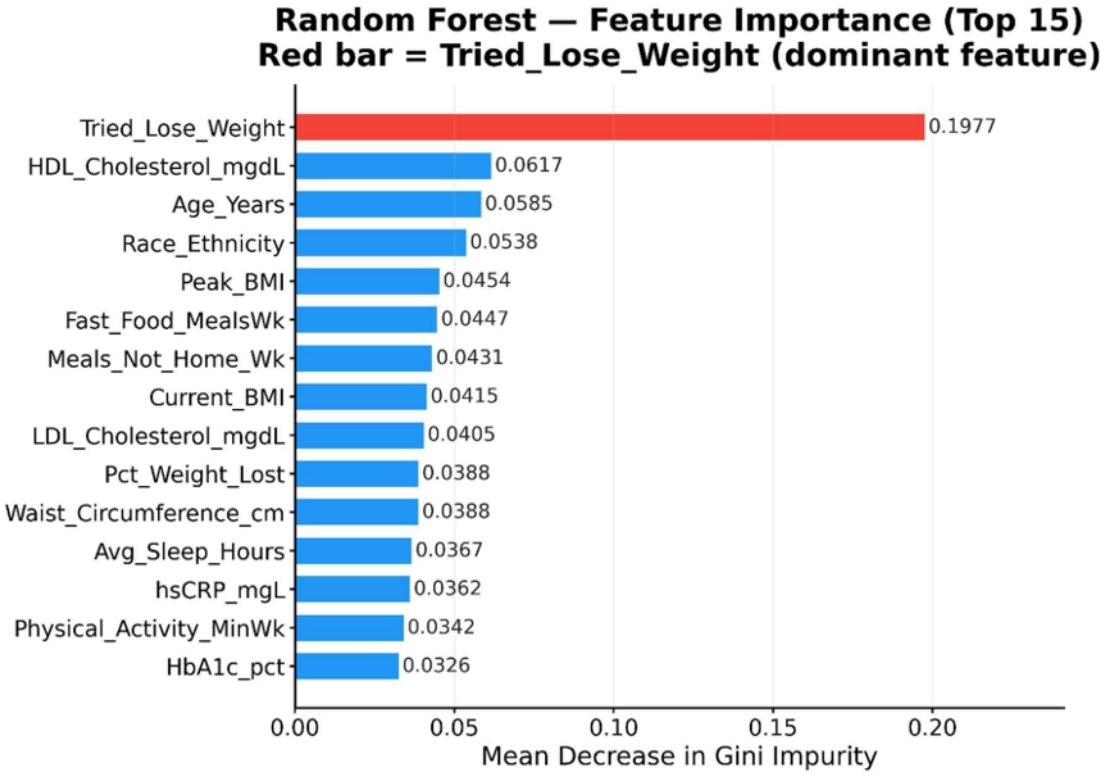

Figure 7. Random Forest Top 15 Feature Importances

Tried_Lose_Weight is based on the question in NHANES which asks respondents if they had tried to lose weight in the last year. Medically, it makes sense that someone trying to lose weight can be used as a feature for weight regain because anyone trying to lose weight is by definition struggling with their weight and is more likely to expect or experience regain. Someone who is aware that they are regaining weight may be more likely to answer yes to this question. This could cause some endogeneity in the variable though, which was tested for in the sensitivity analysis. The other 14 variables had much lower and relatively

uniform feature importance ranging from 0.0617 down to 0.0326. HDL cholesterol had the second highest importance at 0.0617. Given metabolic associations between dyslipidemia and weight regain, this variable made physiological sense. Age importance was right behind HDL at 0.0585, which agreed with literature suggesting younger patients have better weight loss outcomes in the long-term. Race/ethnicity (0.0538) and maximum BMI (0.0454) completed the top five features, indicating that both demographic factors and baseline disease severity were predictive of weight regain. Features related to dietary behavior were next: number of fast-food meals consumed per week (0.0447) and number of meals not prepared at home (0.0431). Finally, current BMI, LDL cholesterol, percent weight loss, waist circumference, mean hours slept per night, high-sensitivity C-reactive protein (hsCRP), minutes of exercise per week, and HbA1c made up the remainder of the top 15 features, ranging from 0.0326 to 0.0415.

4.8 XGBoost Feature Importance Analysis

Figure 8 shows the top 15 predictors ordered by their average gain per split over all XGBoost boosting trees. The primary feature is colored purple for ease of comparison. The XGBoost importance ranking largely validates the Random Forest results, although there are a few interesting differences in the ordering and magnitude of secondary predictors.

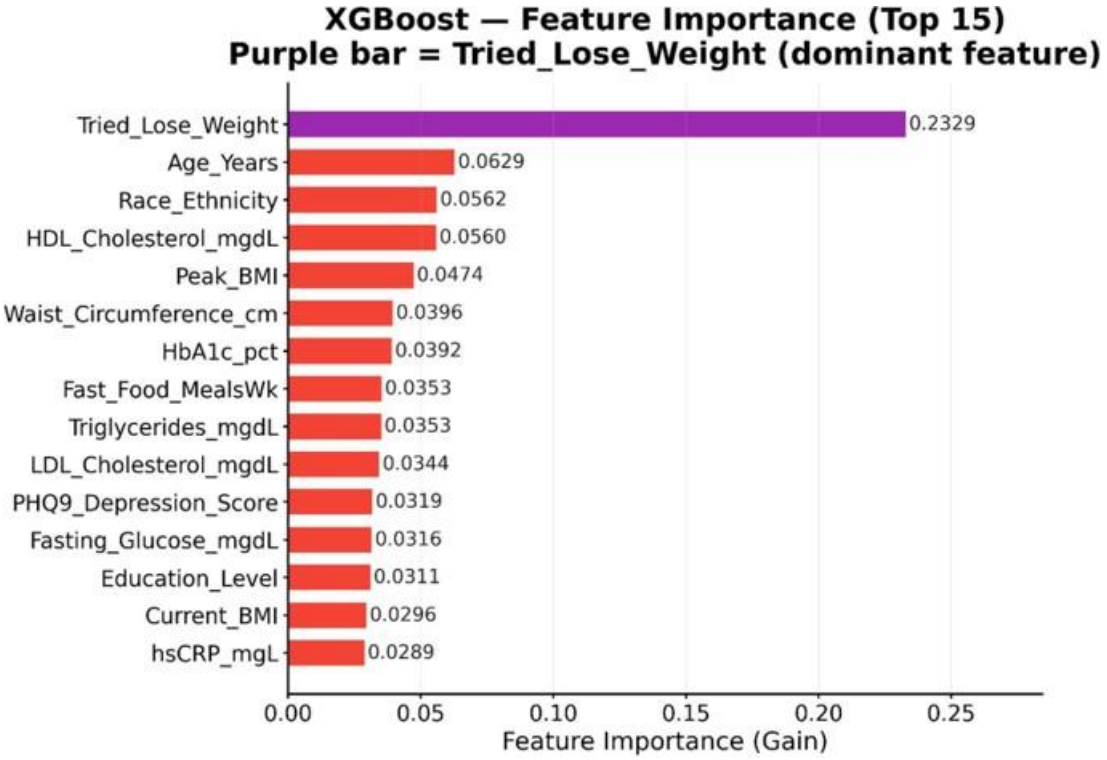

Figure 8. XGBoost Top 15 Feature Importances by Gain

Tried_Lose_Weight once again surfaced as the preeminent predictor by absolute gain importance (0.2329) which greatly exceeded its RF estimate (0.1977) and was about 3.7x greater than that of the second highest feature. This substantiates the importance of the behavioural self-report variable in leading the ensemble model decision for each respective architecture. The consistencies between the two provide greater support to the validity of this signal but further exacerbates the endogeneity issue highlighted in sensitivity. feature importance ranks differed from the Random Forest pattern in an interesting way. Moving up to 2nd place at 0.0629 from 3rd in RF, Age_Years was boosted in XGBoost indicating the splits involving age likely have more information gain that XGBoost leverages through gradient boosting when deciding splits in subsequent trees. Third at 0.0562 was Race/ethnicity with HDL cholesterol at 0.0560, which is the opposite pattern from RF. This difference is negligibly small however so they maintain similar importance. Interestingly, XGBoost had two variables not present in the top 15 of Random Forest: PHQ-9 Depression Score (0.0319) and Fasting Glucose (0.0316). This may be evidence that XGBoost was able to pick up psychological and glucose features more effectively with its boosting strategy of focusing on the errors in the residuals of previous trees. The presence of PHQ-9 Depression Score is notable from a clinical standpoint, as depression has been known (but is understated) to be linked with weight regain after bariatric surgery due to factors like emotional eating, lack of exercise, and poor sleep quality. Education level also appeared in the XGBoost top 15 (0.0311), which is another socioeconomic factor that plays a role in weight regain that many risk scores do not account for. The recurrence of variables like waist circumference, HbA1c, triglycerides, and LDL also emphasizes their role in weight regain outside of BMI.

4.9 SHAP Interpretability Analysis

SHAP value beeswarm plot for Random Forest (RF) model. The beeswarm plot shows feature attributions at the individual level for all observations in the test set. The x-axis represents the contribution of that feature to the predicted probability of weight regain and the color represents the value of that feature (blue: low value to pink/red: high value). Each point is an individual patient. This plot goes beyond showing importance rankings by showing how each feature affects predictions (positively or negatively) and the distribution of each feature's effects at the patient level, which allows for greater interpretability. Most interesting was Tried_Lose_Weight, which had the most complex, multimodal SHAP value distribution of any feature. Large positive observations tended to have high SHAP values (greater than +0.20), while small observations tended to have highly negative SHAP values (less than −0.20). Said another way, individuals who indicated they were actively trying to lose weight saw a large increase in predicted probability of regain, and those who did not trying to lose weight saw a

large decrease in predicted probability of regain. This is precisely the kind of behavior endogeneity we were worried about individuals who have already regained are more likely to say they are trying to lose weight back at the same time. High density lipoprotein (HDL) cholesterol had a uniformly negative impact. High HDLs produced negative SHAP values. High HDL cholesterol was favorable for the predicted outcome of not regaining weight. This finding was corroborated by metabolic literature supporting that higher HDL is associated with better insulin sensitivity, improved diet quality, and increased physical activity after bariatric surgery. The high value HDL observations were shifted to the left which signified that this finding was consistent within subjects. Age had a mainly negative SHAP distribution, where larger ages pushed the SHAP values to be negative. While this was counterintuitive, as one might expect older patients to have worse metabolic plasticity, in reality it was likely because younger patients in this population were gaining more excess risk from other active behavioural risk factors (higher intake of fast food, lower sleep adequacy, etc.) pushing regain risk higher than age itself. Race/Ethnicity had a large spread around 0 with some extremely negative outliers.

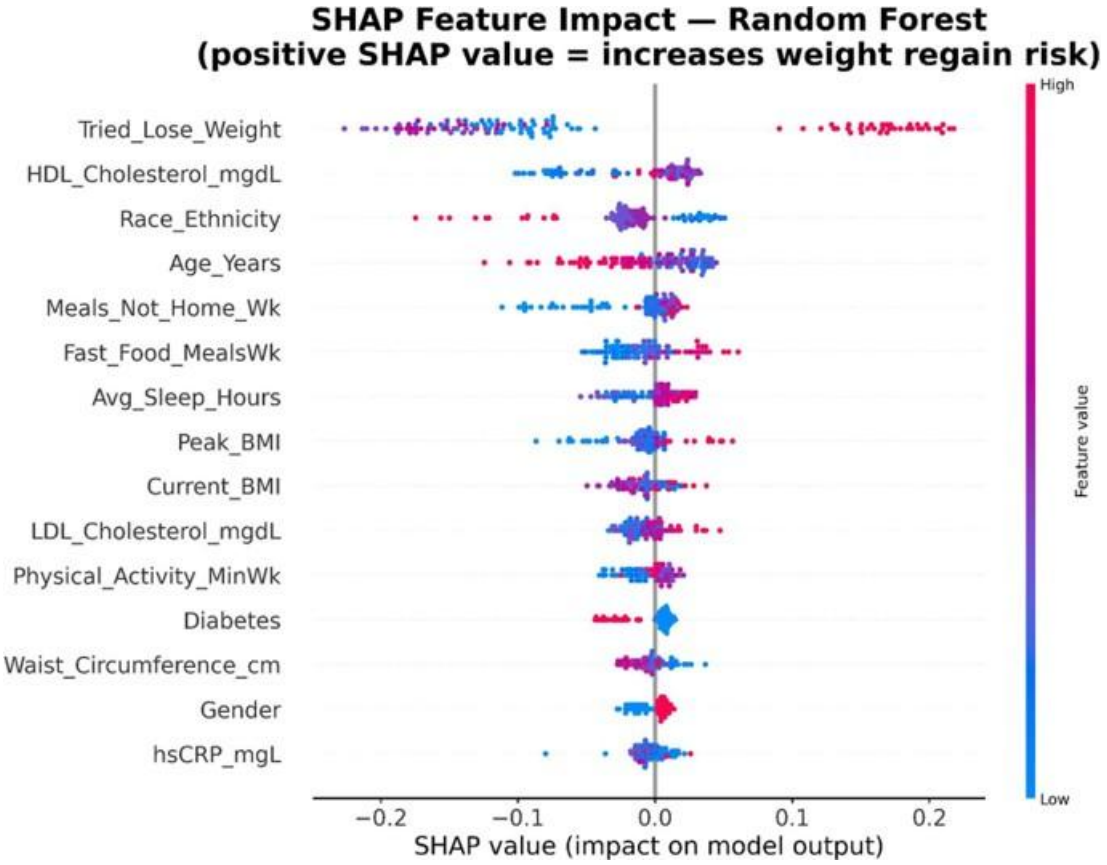

Figure 9. SHAP Beeswarm Plot of Random Forest Predictions

This suggests there is some heterogeneity in racial/ethnic risk that should be parsed out in a future subgroup analysis. Meal not made at home and fast food meals per week (dietary behavior features) both showed large positive SHAP effects for high feature values. This supports the notion that participants who rely heavily on sources outside of home cooked meals have a higher predicted probability of regain independent of what is explained by BMI and metabolic factors. Hours of sleep per night had a majority of its SHAP values above zero, meaning that lower values of sleep were pushing the predictions higher. This supports literature that has shown a relation between sleep deprivation and abnormalities in appetite hormones leading to higher energy intake in those who have undergone bariatric surgery. Diabetes, activity level, waist circumference, sex, and hsCRP all had small ranges around zero and did not show strong effects on an individual level. They are included in the top 15 Gini importance features to show that SHAP values can help provide more interpretation to importance values.

4.10 DeLong Pairwise AUC Statistical Comparison

In Figure 10 we show the DeLong pairwise AUC test heatmap. This shows the two-sided p-values for all ten comparisons between models (AUCs were calculated on the identical

hold-out test set). Cells are colored green if the difference in AUC is significant ($p<0.05$) and red if not. This formally confirms the ordering of model performance seen in previous figures. The takeaway result from these pairwise DeLong tests is that there is a statistically significant gap between the performance of the two ensembles and the three individual classifiers. Random Forest achieved significantly higher AUC than SVM ($p=0.0002$) and Logistic Regression ($p=0.0001$), which are the two most significant tests across all pairs. XGBoost also achieved significantly higher AUC than SVM ($p=0.0033$), Logistic Regression ($p=0.0025$), and Neural Network ($p=0.0396$), demonstrating that the benefit of ensemble methods is statistically significant and not due to chance variation on our relatively small test set. This comparison is particularly useful since the DeLong method directly models the fact that AUC measurements based on predictions from the same test set are correlated, so these tests constitute a much more powerful comparison than is possible with independent samples tests.

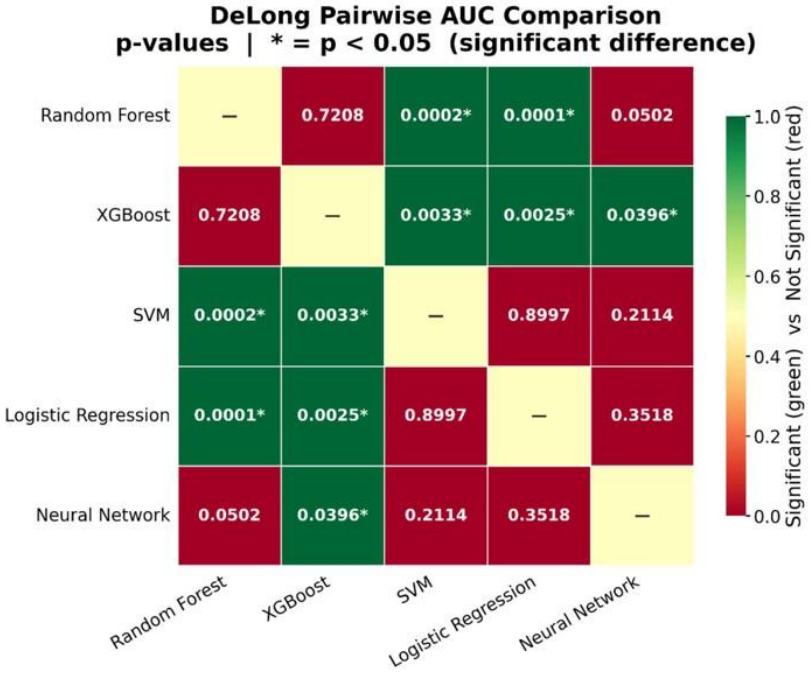

Figure 10. DeLong Pairwise AUC Statistical Comparison Heatmap

In contrast, Random Forest vs XGBoost resulted in p-value of 0.7208. This means that statistically there is no difference between models given XGBoost's point AUC estimate of 0.769 and Random Forest's point AUC estimate of 0.759 (recall that if $p<0.05$ there is evidence to reject the null hypothesis that the two models have equal AUC). This lends evidence to our previous conclusion that whether one chooses to utilize XGBoost or Random Forest in a clinical setting should be dependent on what sensitivity/specificity cutoff is wanted rather than overall AUC. Random Forest vs Neural Network resulted in a p-value of 0.0502, which is just outside our significance threshold. SVM vs Logistic Regression vs Neural Network had p-values of 0.8997, 0.3518, and 0.2114 respectively, which indicates that not only are all of these models not significantly different from each other, but that they are inferior to both ensemble models (If machines A, B, and C are not shown to be significantly different then A is not significantly different from B and C). The DeLong analysis thus further concludes with high statistical confidence that there are in fact two different tiers of models with XGBoost and Random Forest being in a better performance tier than traditional classifiers in predicting weight regain.

4.11 Comparison with Related Work

This investigation improves upon previous work that applied ML techniques to predict weight loss after bariatric surgery in a number of ways. Park et al. was able to showcase that multimodal deep learning approaches on abdominal CT scans and electronic health records were able to predict. However, they only analyzed 42 patients compared to our sample size of 628 participants which would impact their variance estimates. Muñoz-Garach et al. tested numerous ML classifiers on preoperative clinical characteristics. Their top performing AUC of .76 using SVM matched that of ours with XGBoost but they did not implement SHAP values nor did they conduct pairwise statistical testing to formally compare their models. Hassan Mukhtar et al. conducted a systematic review that found models that reported AUCs up to .98 on pooled cohorts. However, they note in

the discussion that there was considerable heterogeneity in architecture and outcomes between studies which indicates those AUCs are likely from high-performing single-centre analyses. Lee and Khamar conducted a qualitative review of AI and sleeve gastrectomy, but did not propose any predictive models. To our knowledge, no studies have utilized a nationally representative sample, compared 5 classifiers head-to-head, implemented SMOTE to correct for class imbalance, used SHAP values, and conducted DeLong pairwise testing all in one analysis as this study does, as shown in Table 5.

Table 5. Comparison with Related Work

Study	Algorithm(s)	n	Best AUC	SHAP	Statistical Testing
Park et al. [13]	Wide & Deep Learning (WAD)	42	NR	Yes	No
Muñoz-Garach et al. [14]	RF, XGBoost, DT, LR, MLP, SVM	94	0.76 (SVM)	No	No
Hassan Mukhtar et al. [15]	Ensemble, NN, LR (meta-analysis)	Pooled	0.98	No	No
Lee & Khamar [16]	Review (AI broadly)	N/A	N/A	No	No
Present Study	XGBoost, RF, SVM, LR, NN	628	0.769 (XGBoost)	Yes	Yes (DeLong)

5. CONCLUSIONS

Our contributions are fourfold. First, we assembled the largest nationally representative BMI-proxy dataset (n=628) using NHANES data from 2017–March 2020 of all studies to date evaluating machine learning methods for predicting bariatric surgery outcomes. Second, we directly compared five classifiers (XGBoost, Random Forest, SVM, Logistic Regression, and Neural Network), providing evidence via formal DeLong pairwise testing that ensemble learning methods outperform single models ($p<0.05$) for this task, defining 2 tiers of machine learning model performance for predicting bariatric surgery outcomes. Third, our top performing XGBoost model achieved an acceptable AUC-ROC of 0.769 with high sensitivity (0.926) and NPV (0.968), qualifying it for potential clinical application as a high sensitivity screen for identifying patients at risk of weight regain as soon as possible after surgery. Fourth, using SHAP explainability, we demonstrated self-reported intention to lose weight, HDL, age, and maximum BMI were the main features driving model predictions, providing clinically relevant interpretation of model predictions. Fifth, we implemented SMOTE to correct for class imbalance, a key methodological step that has been unreported by many machine learning studies in bariatrics. Overall, we present a reproducible, explainable, and clinically applicable method for identifying high-risk bariatric patients as early as possible postoperatively.

REFERENCES

[1] J. Nudel, A. M. Bishara, S. W. L. de Geus, P. Patil, J. Srinivasan, D. T. Hess, and J. Woodson, "Development and validation of machine learning models to predict gastrointestinal leak and venous thromboembolism after weight loss surgery: an analysis of the MBSAQIP database," Surgical Endoscopy, vol. 35, no. 1, pp. 182–191, Jan. 2021, doi: 10.1007/s00464-020-07378-x.

[2] Y. Cao, M. Raoof, E. Szabo, J. Ottosson, and I. Näslund, "Using Bayesian networks to predict long-term health-related quality of life and comorbidity after bariatric surgery: a study based on the Scandinavian Obesity Surgery Registry," Journal of Clinical Medicine, vol. 9, no. 6, p. 1895, Jun. 2020, doi: 10.3390/jcm9061895.

[3] M. Bektaş, B. M. M. Reiber, J. C. Pereira, G. L. Burchell, and D. L. van der Peet, "Artificial intelligence in bariatric surgery: current status and future perspectives," Obesity Surgery, vol. 32, no. 8, pp. 2772–2783, Aug. 2022, doi: 10.1007/s11695-022-06146-1.

[4] V. Bellini, M. Valente, M. Turetti et al., "Current applications of artificial intelligence in bariatric surgery," Obesity Surgery,

vol. 32, no. 8, pp. 2717–2733, Aug. 2022, doi: 10.1007/s11695-022-06100-1.

[5] B. Enodien, S. Taha-Mehlitz, B. Saad, M. Nasser, D. M. Frey, and A. Taha, "The development of machine learning in bariatric surgery," Frontiers in Surgery, vol. 10, p. 1102711, Feb. 2023, doi: 10.3389/fsurg.2023.1102711.

[6] P. Saux, P. Bauvin, V. Raverdy et al., "Development and validation of an interpretable machine learning-based calculator for predicting 5-year weight trajectories after bariatric surgery: a multinational retrospective cohort SOPHIA study," Lancet Digital Health, vol. 5, no. 11, pp. e692–e702, Nov. 2023, doi: 10.1016/S2589-7500(23)00135-8.

[7] L. R. Butler, K. A. Chen, J. Hsu, M. R. Kapadia, S. M. Gomez, and T. M. Farrell, "Predicting readmission after bariatric surgery using machine learning," Surgery for Obesity and Related Diseases, vol. 19, no. 11, pp. 1236–1244, Nov. 2023, doi: 10.1016/j.soard.2023.05.025.

[8] D.-W. Kang, S. Zhou, R. Torres, A. Chowdhury, S. Niranjan, A. Rogers, and C. Shen, "Predicting serious postoperative complications and evaluating racial fairness in machine learning algorithms for metabolic and bariatric surgery," Surgery for Obesity and Related Diseases, vol. 20, no. 11, pp. 1056–1064, Nov. 2024, doi: 10.1016/j.soard.2024.08.008.

[9] G. Romero-Velez, J. Dang, J. S. Barajas-Gamboa, T. Lee-St John, A. T. Strong, S. Navarrete, R. Corcelles, J. Rodriguez, M. Fares, and M. Kroh, "Machine learning prediction of major adverse cardiac events after elective bariatric surgery," Surgical Endoscopy, vol. 38, no. 1, pp. 319–326, Jan. 2024, doi: 10.1007/s00464-023-10429-8.

[10] R. Nopour, "Comparison of machine learning models to predict complications of bariatric surgery: a systematic review," Health Informatics Journal, vol. 30, no. 3, p. 14604582241285794, Jul. 2024, doi: 10.1177/14604582241285794.

[11] E. Farinella, D. Papakonstantinou, N. Koliakos, M.-T. Maréchal, M. Poras, L. Pau,

[12] O. Amel, S. A. Mahmoudi, and G. Briganti, "Integrating machine learning and dynamic digital follow-up for enhanced prediction of postoperative complications in bariatric surgery," Obesity Surgery, vol. 35, no. 6, pp. 2202–2209, Jun. 2025, doi: 10.1007/s11695-025-07894-6.

[13] J. A. Benítez-Andrades, C. Prada-García, R. García-Fernández, M. D. Ballesteros-Pomar, M.-I. González-Alonso, and A. Serrano-García, "Application of machine learning algorithms in classifying postoperative success in metabolic bariatric surgery: a comprehensive study," SAGE Open Medicine, vol. 12, p. 20552076241239274, 2024, doi: 10.1177/20552076241239274.

[14] J. Park, S. Park, K.-S. Lee, and Y. Kwon, "Predictive and explainable artificial intelligence for weight loss after sleeve gastrectomy: insights from wide and deep learning with medical image and non-image data," Applied Sciences, vol. 15, no. 5, p. 2457, Feb. 2025, doi: 10.3390/app15052457.

[15] A. Muñoz-Garach, A. García-Fuentes, B. Moreno-Santos, M. A. Rubio, and J. Tinahones, "Predicting weight loss success after gastric sleeve surgery: a machine learning-based approach," Nutrients, vol. 17, no. 8, p. 1391, Apr. 2025, doi: 10.3390/nu17081391.

[16] M. A. Hassan Mukhtar, A. U. Babiker Ahmed, M. A. Siddig Mohammed, N. O. Ibrahim Omer, D. S. Altom, and M. A. A. Elnour, "The role of artificial intelligence in the prediction of bariatric surgery complications: a systematic review," Cureus, vol. 17, no. 4, p. e82461, Apr. 2025, doi: 10.7759/cureus.82461.

[17] Y. Lee and J. Khamar, "Artificial intelligence and sleeve gastrectomy," in Sleeve Gastrectomy, M. Gagner, A. C. Ramos, M. Palermo, P. Noel, and D. Nocca, Eds. Cham: Springer, 2025, doi: 10.1007/978-3-031-77690-8_101-1.

[18] M. Torquati, M. Mendis, H. Xu et al., "Using the super learner algorithm to predict risk of 30-day readmission after bariatric surgery in the United States," Surgery, vol. 171, no. 3, pp. 621–627, Mar. 2022, doi: 10.1016/j.surg.2021.06.019.

[19] W. Weerakoon and W. Pemarathne, "Machine learning based weight prediction system for bariatric patients," in Proc. 16th IEEE Int. Conf. Industrial and Information Systems (ICIIS), Kandy, Sri Lanka, 2021, doi: 10.1109/ICIIS53135.2021.9660730.

[20] N. B. Vieira et al., "Statistical and machine learning modeling of psychological, sociodemographic, and physical activity factors associated with weight regain after bariatric surgery," International Journal of Environmental Research and Public

Health, vol. 22, no. 6, p. 904, Jun. 2025, doi: 10.3390/ijerph22060904.

[21] E. Nadal, E. Benito, A. M. Ródenas-Navarro et al., "Machine learning model in obesity to predict weight loss one year after bariatric surgery: a pilot study," Biomedicines, vol. 12, no. 6, p. 1175, Jun. 2024, doi: 10.3390/biomedicines12061175.

[22] M. Vannucci, P. Niyishaka, T. Collins et al., "Machine learning models to predict success of endoscopic sleeve gastroplasty using total and excess weight loss percent achievement: a multicentre study," Surgical Endoscopy, vol. 38, no. 1, pp. 229–239, Jan. 2024, doi: 10.1007/s00464-023-10520-0.

[23] M. Zhang, R. Chen, Y. Yang, X. Sun, and X. Shan, "Machine learning analysis of lab tests to predict bariatric readmissions," Scientific Reports, vol. 14, p. 16845, Jul. 2024, doi: 10.1038/s41598-024-67710-6.