Surgical Endoscopy (2021) 35:182–191
https://doi.org/10.1007/s00464-020-07378-x

Development and validation of machine learning models to predict gastrointestinal leak and venous thromboembolism after weight loss surgery: an analysis of the MBSAQIP database

Jacob Nudel[1,2] · Andrew M. Bishara[3,4] · Susanna W. L. de Geus[1] · Prasad Patil[5] · Jayakanth Srinivasan[2] · Donald T. Hess[1] · Jonathan Woodson[2]

Received: 23 April 2019 / Accepted: 7 January 2020 / Published online: 17 January 2020

Abstract

Background Postoperative gastrointestinal leak and venous thromboembolism (VTE) are devastating complications of bariatric surgery. The performance of currently available predictive models for these complications remains wanting, while machine learning has shown promise to improve on traditional modeling approaches. The purpose of this study was to compare the ability of two machine learning strategies, artificial neural networks (ANNs), and gradient boosting machines (XGBs) to conventional models using logistic regression (LR) in predicting leak and VTE after bariatric surgery.

Methods ANN, XGB, and LR prediction models for leak and VTE among adults undergoing initial elective weight loss surgery were trained and validated using preoperative data from 2015 to 2017 from Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program database. Data were randomly split into training, validation, and testing populations. Model performance was measured by the area under the receiver operating characteristic curve (AUC) on the testing data for each model.

Results The study cohort contained 436,807 patients. The incidences of leak and VTE were 0.70% and 0.46%. ANN (AUC 0.75, 95% CI 0.73–0.78) was the best-performing model for predicting leak, followed by XGB (AUC 0.70, 95% CI 0.68–0.72) and then LR (AUC 0.63, 95% CI 0.61–0.65, $p < 0.001$ for all comparisons). In detecting VTE, ANN, and XGB, LR achieved similar AUCs of 0.65 (95% CI 0.63–0.68), 0.67 (95% CI 0.64–0.70), and 0.64 (95% CI 0.61–0.66), respectively; the performance difference between XGB and LR was statistically significant ($p = 0.001$).

Conclusions ANN and XGB outperformed traditional LR in predicting leak. These results suggest that ML has the potential to improve risk stratification for bariatric surgery, especially as techniques to extract more granular data from medical records improve. Further studies investigating the merits of machine learning to improve patient selection and risk management in bariatric surgery are warranted.

Keywords Bariatric surgery · Postoperative complications · Anastomotic leak · Venous thromboembolism · Machine learning · Deep learning

Obesity and associated metabolic diseases constitute a major public health threat for which bariatric surgery is a highly effective intervention [1]. Laparoscopic weight loss surgery (WLS) is safe relative to other elective general

Presented at the American College of Surgeons Quality and Safety conference, July 2019

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00464-020-07378-x) contains supplementary material, which is available to authorized users.

✉ Jonathan Woodson
jwoodson@bu.edu

1 Department of Surgery, Boston University School of Medicine, Boston, MA, USA

2 Institute for Health System Innovation and Policy, Boston University, 601, 656 Beacon Street, Boston, MA 02215, USA

3 Department of Anesthesia, University of California, San Francisco, San Francisco, CA, USA

4 Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, CA, USA

5 Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA

surgical procedures [2], but complications can be morbid and expensive [3]. Safety concerns among both patients [4] and providers [5] help explain why WLS is under utilized relative to clinical needs [6]. Stratification of risk for postoperative complications can guide patient selection, inform referral practices and patient counseling, and identify high-risk patients for monitoring and intervention.

Gastrointestinal leak occurs in less than one percent of WLS cases [7] but is associated with other complications, readmission, reoperation, death [8], and increased cost [9]. Obese patients are at high risk for deep vein thrombosis [10, 11] and American Society for Metabolic and Bariatric Surgery guidelines recommend routine thromboprophylaxis [12]. Nevertheless, venous thromboembolism (VTE) remains a leading cause of morbidity and mortality in this population [13, 14] and optimizing thromboprophylaxis strategies remains an area of considerable interest [13, 15, 16]. Prior risk models for leak and VTE achieve modest results [14, 17]. For example, BariClot is a VTE risk assessment tool based on logistic regression (LR) that was developed and validated using the Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program (MBSAQIP) registry. Though it achieved an area under the receiver operating characteristic curve (AUC) of just 0.60, it outperformed two previously published models [14, 18, 19].

Machine learning (ML), a branch of artificial intelligence, is the study of computer algorithms that extract information from data without explicit instructions from humans. ML does not refer to a specific mathematical approach, but to a broad array of statistical models. These are generally related in their flexibility and capacity to distinguish subtle, nonlinear patterns in data that are often not accessible to traditional approaches like LR [20]. ML models have recently outperformed LR in preoperative risk stratification using National Surgical Quality Improvement Program data [21, 22].

Artificial neural networks (ANNs) and gradient boosting machines (XGBs) are powerful classes of ML models that perform well in medical risk prediction using tabular data [23, 24]. A simple ANN is a stack of layered functions with each layer containing a matrix of weights. Data pass through the stack with the output of one layer used as the input to the next, ultimately transforming the data into model outputs. Training involves repeatedly adjusting the weights to gradually match model to target outputs [25]. XGB is a ML algorithm in which a series of decision models are iteratively constructed, tested, and adjusted to correct outputs, ultimately resulting in a decision tree algorithm optimized for a regression or classification task [26].

The aim of this study was to develop and validate preoperative ANN and XGB risk models for gastrointestinal leak and VTE among WLS patients and compare their performance against traditional models.

Methods

Data source and study population

All available MBSAQIP data from 2015 to 2017 were used. This national registry contains patient-level variables characterizing preoperative risk factors and 30-day postoperative outcomes. In 2017, 832 accredited bariatric centers contributed over 200,000 cases to the registry [27]. The study population included patients aged 18–79 with no prior foregut or bariatric surgery who underwent elective laparoscopic gastric bypass (CPT 43644 or 43645) or laparoscopic sleeve gastrectomy (CPT 43775). We excluded patients with no information on height and weight or Body Mass Index (BMI) given the fundamental importance of this information to the study interventions. This study was approved by the Boston Medical Center Institutional Review Board under a pre-existing protocol for research on MBSAQIP data.

Outcomes

Outcomes of interest were gastrointestinal leak and VTE. Each was defined as a composite endpoint of 30-day outcomes variables in MBSAQIP. Leak was defined as postoperative organ space infection, presence of a surgical drain for more than 30 days, or leak as the suspected reason for any readmission, reintervention, or reoperation [7]. VTE was defined as anticoagulation therapy for imaging-confirmed deep vein thrombosis (DVT) or pulmonary embolism (PE) or readmission, reintervention, reoperation, or death with DVT or PE as the suspected cause [14].

Predictive models

For each outcome of interest, we randomly split the data into training, validation, and testing populations comprising 50%, 25%, and 25% of the study cohort respectively. To account for imbalanced data, we oversampled positive cases to a ratio of 0.5 in the training set using the imbalanced learn Python library [28, 29]. Positive and negative cases were split separately to ensure equitable distribution of positive cases in the training, validation, and testing sets.

Predictive models used all clinical variables that could be reasonably ascertained the day prior to surgery (Table 1). To permit valid comparisons of model performance, all models used all available input variables to generate predictions. Some features were calculated or consolidated from MBSAQIP variables (Table 1). Continuous

Table 1 Input variables and outcomes among 436,807 patients undergoing elective laparoscopic gastric bypass or sleeve gastrectomy

Input variable	
Sex, *n* (%)	
Female	346,559 (79.3)
Male	90,248 (20.7)
Race, *n* (%)	
American Indian or Alaska Native	1745 (0.4)
Asian	2138 (0.5)
Black or African American	77,050 (17.6)
Native Hawaiian or Other Pacific Islander	1222 (0.3)
Unknown/not reported	35,193 (8.1)
White	319,459 (73.1)
Hispanic ethnicity, *n* (%)	
No	340,748 (78.0)
Unknown	41,535 (9.5)
Yes	54,524 (12.5)
Procedure, *n* (%)	
Gastric bypass	121,528 (27.8)
Sleeve gastrectomy	315,279 (72.2)
Gastroesophageal reflux disease, *n* (%)	
No	301,408 (69.0)
Yes	135,399 (31.0)
Limited ambulation, *n* (%)	
No	429,440 (98.3)
Yes	7367 (1.7)
Vein thrombosis requiring therapy, *n* (%)	
No	429,833 (98.4)
Yes	6974 (1.6)
History of myocardial infarction, *n* (%)	
No	431,143 (98.7)
Yes	5664 (1.3)
Previous PCI or angioplasty, *n* (%)	
No	427,889 (98.0)
Yes	8918 (2.0)
Previous cardiac surgery, *n* (%)	
No	431,964 (98.9)
Yes	4843 (1.1)
Hypertension requiring medication, *n* (%)	
No	224,663 (51.4)
Yes	212,144 (48.6)
Number of anti-Hypertensive medications, *n* (%)	
0	159,267 (36.5)
1	94,885 (21.7)
2	72,381 (16.6)
3+	110,274 (25.2)
Hyperlipidemia, *n* (%)	
No	331,523 (75.9)
Yes	105,284 (24.1)
Venous stasis, *n* (%)	
No	432,278 (99.0)
Yes	4529 (1.0)

Table 1 (continued)

Input variable	
Dialysis requirement, *n* (%)	
No	435,460 (99.7)
Yes	1347 (0.3)
Renal insufficiency, *n* (%)	
No	433,915 (99.3)
Yes	2892 (0.7)
Preoperative therapeutic anticoagulation, *n* (%)	
No	425,520 (97.4)
Yes	11,287 (2.6)
Diabetes, *n* (%)	
Insulin dependent	38,102 (8.7)
No	320,820 (73.4)
NonInsulin dependent	77,885 (17.8)
Smoker, *n* (%)	
No	399,223 (91.4)
Yes	37,584 (8.6)
Functional status, *n* (%)	
Independent	432,220 (98.9)
Partially dependent	2833 (0.6)
Totally dependent	1754 (0.4)
Chronic obstructive pulmonary disease, *n* (%)	
No	429,313 (98.3)
Yes	7494 (1.7)
Oxygen dependent, *n* (%)	
No	433,635 (99.3)
Yes	3172 (0.7)
History of pulmonary embolism, *n* (%)	
No	431,748 (98.8)
Yes	5059 (1.2)
Sleep apnea, *n* (%)	
No	269,762 (61.8)
Yes	167,045 (38.2)
Chronic steroids, *n* (%)	
No	429,452 (98.3)
Yes	7355 (1.7)
Presence and timing of placement of IVCF, *n* (%)[a]	
Placed in anticipation of surgery	2243 (0.5)
Pre-existing	978 (0.2)
No	433,539 (99.3)
Unknown	47 (0.0)
American Society of Anesthesiology Class, *n* (%)	
1—No disturb	1476 (0.3)
2—Mild disturb	97,939 (22.4)
3—Severe disturb	319,773 (73.2)
4—Life threat	15,571 (3.6)
5—Moribund	40 (0.0)
Unknown	2008 (0.5)
Training level of first assistant, *n* (%)	
Attending—other	24,369 (5.6)
Attending—weight loss surgeon	65,444 (15.0)

Table 1 (continued)

Input variable	
Minimally invasive surgery fellow	38,613 (8.8)
None (no assist or scrub tech/RN only)	63,273 (14.5)
Physician assistant/nurse practitioner/registered nurse	166,222 (38.1)
Resident (PGY 1–5+)	78,886 (18.1)
Year of operation, *n* (%)	
2015	131,926 (30.2)
2016	146,614 (33.6)
2017	158,267 (36.2)
Height in centimeters, mean (sd)	166.7 (9.2)
Consolidated preoperative BMI, mean (sd)[b]	45.4 (8.0)
Change in BMI in the year prior to surgery, mean (sd)[c]	− 2.0 (2.3)
Weight in kilograms, mean (sd)[d]	126.7 (26.8)
Age in years, mean (sd)[e]	44.7 (12.0)
Preoperative albumin level, mean (sd)[f]	4.1 (0.4)
Preoperative hematocrit level, mean (sd)[g]	40.9 (3.8)
Operative duration (minutes)[h]	85.8 (47.1)
Outcomes	
Gastrointestinal leak, *n* (%)	
No	433,739 (99.3)
Yes	3068 (0.7)
Venous thromboembolism, *n* (%)	
No	434,795 (99.5)
Yes	2012 (0.5)

BMI body mass index, *PCI* percutaneous coronary intervention, *IVCF* inferior vena cava filter

[a]The presence and timing of placement of preoperative inferior vena cava filters were consolidated into one variable

[b]In the event that preoperative BMI was available but maximum BMI for the preceding year was not, the most recent BMI was assumed to be the maximum BMI ($n=27{,}862$); when preoperative BMI was not available, it was set equal to the maximum ($n=2268$)

[c]A continuous variable representing the difference between the maximum BMI and the preoperative BMI was computed. 27,862 missing

[d]Back calculated from height and consolidated BMI

[e]The 2015 MBSAQIP PUF reports ages as digits, whereas the 2016 and 2016 PUFs report ages to the hundredth decimal place. To avoid losing information from the latter cohorts, we reassigned each 2015 patient a randomly selected age from a uniform distribution within the appropriate year

[f]114,343 missing

[g]44,969 missing

[h]Used only in BariClot calculation

variables were zero centered and scaled to unit variance. Methods for handling missing and incomplete data are described in Table 1. Wherever possible, missing continuous variables were set to the training population mean. Missing categorical variables were assigned to a unique, unknown category.

ANN and XGB were compared to LR for prediction of both VTE and leak. Our ANN, XGB, and LR models were compared to BariClot for prediction of VTE. Our models computed the probability of an outcome for each patient, while BariClot generated a risk score [14]. All predictive models were implemented in Python 3.6 [30, 31] using the Anaconda Distribution [32] with extensive use of the Pandas [33] and NumPy [34] libraries. We followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis reporting guidelines [35]. All code used for preprocessing data and building predictive models is open sourced.

ANN models were implemented in Pytorch [36] with code adapted from open sources [37–39]. ANN architecture consisted of two layers with rectified linear units applied after each layer. We selected a relatively simple architecture because initial experiments with more complex architectures increased computational demand without a notable increase in predictive power. Categorical variables were encoded as neural embeddings [40]. Batch normalization was applied between layers [41]. Early stopping [42] and random dropout [43] were employed to avoid over-fitting training data [23]. Training was terminated when the ANN achieved peak performance on the validation data. XGB was implemented in XGBoost using default hyperparameters [26]. LR was implemented in statsmodels [44].

Statistical comparison of model performance

Model performance was measured by computing the AUC generated by each model on the test set for each outcome. The Delong test [45] with threshold of 0.05 was used to statistically compare AUCs generated by each predictive model. AUC confidence intervals were obtained using the Delong procedure. Bootstrapping was used to find confidence intervals for other model performance measures including comparison of partial AUCs. The pROC package [46] with RStudio [47] and R version 3.5.2 [48] was used for all model performance calculations. Plots were made with ggplot2 [49].

Descriptive statistics were computed in using the tableone Python library [50]. Training, validation, and test populations were compared using one-way ANOVA and chi-square tests for continuous and categorical variables, respectively.

Results

The study cohort contained 436,807 patients of whom 3068 (0.070%) developed leak and 2012 (0.046%) suffered VTE (Supplementary Fig. 1). Characteristics of the cohort are shown in Table 1. The training, validation, and testing sets for both gastrointestinal leak and VTE had 218,403, 109,202, and 109,202 patients, respectively. There were no clinically meaningful differences in patient characteristics between training, validation, and test sets, although there were some statistically significant differences (Tables 1 and 2 in the Supplement).

Figure 1 shows model performance for prediction of leak. ANN was the best-performing model with an AUC of 0.75 (95% CI 0.73–0.78). ANN outperformed XGB ($p < 0.001$), which also performed well, achieving an AUC of 0.70 (95% CI 0.68–0.72). Both ANN and XGB significantly outperformed LR ($p < 0.001$ for each comparison), which achieved an AUC of 0.63 (95% CI 0.61–0.65).

ANN achieved a partial AUC of 0.05 under the portion of the ROC with specificity greater than 90%, outperforming both XGB (partial AUC 0.03, $p < 0.001$) and LR (partial AUC 0.01, $p < 0.001$). With the specificity threshold held as close as possible to 0.975, ANN achieved a sensitivity of 0.493 (95% CI 0.458–0.529), a positive predicative value

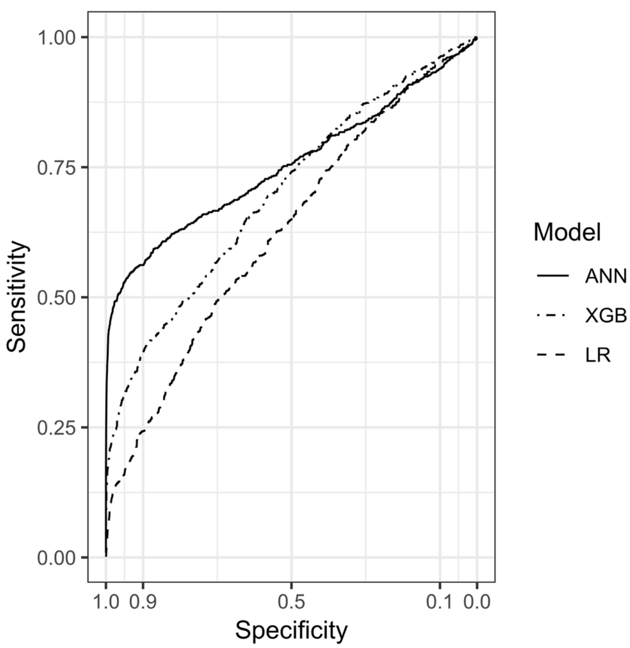

Fig. 1 Receiver Operating Characteristic Curves for Predicting Gastrointestinal Leak. *ANN* artificial neural network, *XGB* gradient boosting machine, *LR* logistic regression

(PPV) of 0.122 (95% CI 0.114–0.131), and outperformed XGB and LR at the same threshold (Table 2). Of the 767 patients in the testing set who went on to suffer postoperative leaks, ANN would have identified 378 at the 0.975 specificity threshold, while XGB and LR would have identified 184 and 103, respectively.

Model performance for prediction of VTE is summarized in Fig. 2. ANN, XGB, and LR achieved similar AUCs of 0.65 (95% CI 0.63–0.68), 0.67 (95% CI 0.64–0.70), and 0.64 (95% CI 0.61–0.66), respectively. XGB outperformed LR ($p = 0.001$) but there were no other statistically significant differences between models. ANN, XGB, and LR outperformed BariClot ($p < 0.001$ for all three comparisons), which achieved an AUC of 0.56 (95% CI 0.54–0.59). At the 0.975 specificity threshold, confusion matrix metrics of the ANN, XGB, and LR models were generally comparable to one another and superior to BariClot (Table 3).

All models used all input variables in prediction. The relative importance of predictive variables in XGB models for both outcomes are shown in Figs. 3 and 4. XGB identified age, height and weight-related measures, hematocrit, albumin, and assistant training level as important predictors for both leak and VTE. History of DVT was among the most important factors in predicting VTE, but not leak (Figs. 3 and 4). Odds ratios for predictive variables used by logistic regression models are listed in the Supplement Tables 3 and 4.

Discussion

This study demonstrates the potential utility of applying ML methods for preoperative risk assessment in bariatric surgery. For predicting leak, ANN and XGB outperformed LR, which performed very similarly to a previously reported LR model [17]. In our study, the potential clinical benefits of ML are most apparent when evaluating our leak models at high specificity, where ANN and XGB performed particularly well and could prove useful in preoperative screening. At 97.5% specificity, ANN predicted several-fold more leaks than LR and achieved a PPV over 10%. Among patients with a 10% probability of leak, the benefits of weight loss surgery are unlikely to outweigh the risks. These results suggest ML

Table 2 Performance characteristics of the artificial neural network (ANN), gradient boosting machine (XGB), and logistic regression (LR) models for predicting gastrointestinal leak at the 97.5% specificity threshold

	Sensitivity, median (95% CI)	Specificity, median (95% CI)	PPV, median (95% CI)
Model			
ANN	0.493 (0.458–0.529)	0.975 (0.974–0.976	0.122 (0.114–0.131)
XGB	0.24 (0.209–0.270)	0.975 (0.974–0.976)	0.063 (0.056–0.071)
LR	0.134 (0.111–0.159)	0.975 (0.974–0.976)	0.037 (0.030–0.043)

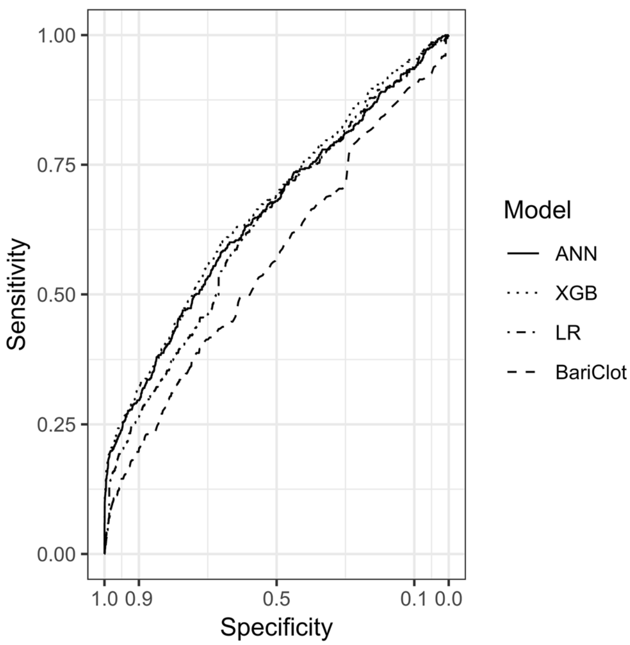

Fig. 2 Receiver operating characteristic curves for predicting venous thromboembolism. *ANN* artificial neural network, *XGB* gradient boosting machine, *LR* logistic regression

methods can offer clinically meaningful improvements in risk stratification, even for uncommon events that are difficult to predict using any statistical method.

In the context of VTE, ANN and XGB perform similarly to LR, with XGB achieving a small but statistically significant advantage. All three of our models outperformed BariClot even though BariClot employs intra-operative information in prediction, likely because BariClot was trained on less data than our models. Recent contributions to the literature on VTE risk after weight loss surgery use a wider range of variables and incorporate patient data from perioperative, intra-operative, and postoperative time points [13, 14, 16]. Our VTE risk models are less predictive than our leak models. This may be because widespread thromboprophylaxis among patients in MBSAQIP dampens the statistical signals available to VTE models.

These results contribute to an emerging literature describing ML for medical risk assessment. ML techniques have recently been applied to tabular data to predict a variety of outcomes including delirium [24] and pediatric emergency department triage [23] with good results. However, ML does not always outperform traditional LR. For example, ML outperformed LR in just one of two recent, rigorous efforts to predict heart failure readmissions, likely due to differences between the data sets used by each team [51–53]. Our results fit the general pattern that no single predictive modeling technique consistently prevails.

Several limitations apply. First, outcomes of interest that occur beyond 30 days or for which patients do not present to the index institution may be missed [54]. However, the effect of increasing the incidence of outcomes in a test population on model performance is unclear, and may actually boost performance. Second, feature selection is limited to the specific variables and level of detail available in MBSAQIP. It is not clear that models developed using narrowly scoped, highly structured data will perform well outside of this context [20, 55]. Nevertheless, our results indicate that ML techniques may provide significant performance gains against LR. ANNs are especially powerful in learning from unstructured and multimodal data. Thus, we suspect access to a wider set of features would have improved the predictive performance of all of our models and of ANNs in particular. Additionally, pretrained ANNs can be adapted to new data in a process called transfer learning. In this fashion, the insights gained through training in large administrative datasets can be harnessed to build high-performing models in specific clinical contexts with relatively small numbers of observations that can be collected on the scale of single institutions [56]. Third, we do not have sufficient data to externally validate our models. ANN and XGB were somewhat overfitted to the training data, but all three of our models performed similarly in the validation and testing data, confirming internal validity (Supplementary Table 5). Fourth, several variables, including the precise age of all patients in the 2015 cohort, were missing in a nontrivial number of cases. However, we split the data to equally distribute the missing data among the training, validation, and testing cohorts, and model performance should therefore account for bias introduced in imputation.

Our ML models are also limited in terms of usability. They employ more variables than clinicians can reasonably input at the point of care. Their utility will depend on assistive software that marries innovation in clinical data

Table 3 Performance characteristics of the artificial neural network (ANN), gradient boosting machine (XGB), logistic regression (LR), and BariClot models for predicting venous thromboembolism at the 97.5% specificity threshold

	Sensitivity, median (95% CI)	Specificity, median (95% CI)	PPV, median (95% CI)
Venous thromboembolism			
ANN	0.203 (0.169–0.239)	0.975 (0.974–0.976)	0.036 (0.03–0.042)
XGB	0.211 (0.175–0.247)	0.975 (0.974–0.976)	0.038 (0.031–0.044)
LR	0.159 (0.127–0.191)	0.975 (0.974–0.976)	0.029 (0.023–0.034)
BariClot	0.101 (0.076–0.127)	0.975 (0.974–0.976)	0.018 (0.014–0.023)

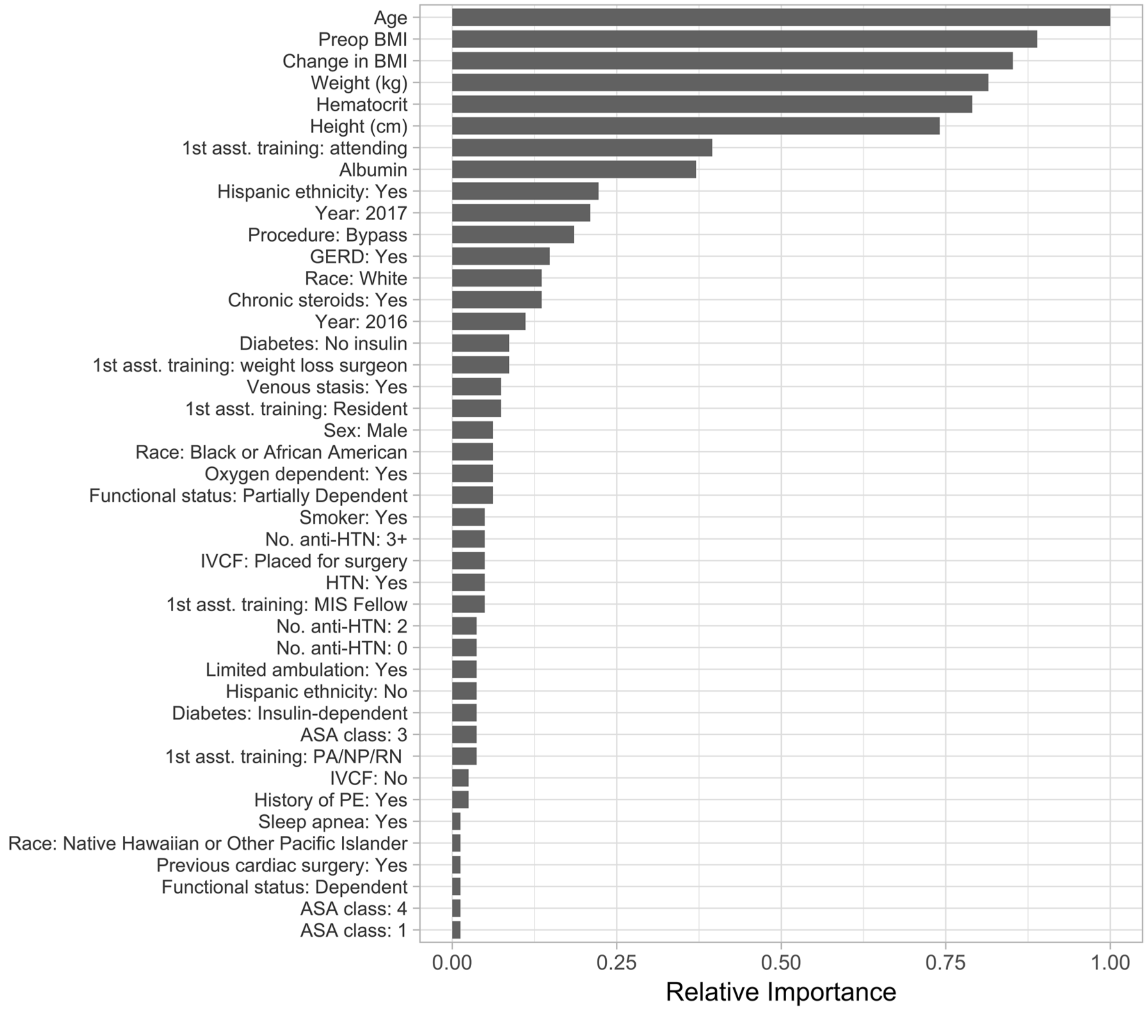

Fig. 3 Relative importance of each predictive variable in the gradient boosting machine model for predicting gastrointestinal leak. Relative performance quantifies the relative contribution of each variable to minimizing the error of the gradient boosting model. The measure is scaled from zero to one against the most important predictor [24]. Relationships between importance and outcomes are nonlinear and cannot be interpreted directionally with respect to their influence on outcomes, nor can they be used to generate cutoff or threshold values. *BMI* body mass index, *DVT* deep vein thrombosis, *MIS* minimally invasive surgery, *PE* pulmonary embolism, *HTN* hypertension, *IVCF* inferior vena cava filter, *PCI* percutaneous coronary intervention, *GERD* gastroesophageal reflux disease, *ASA* American Society of Anesthesiology, *COPD* chronic obstructive pulmonary disease, *HLD* hyperlipidemia, *MI* myocardial infarction

management to user interface design [20, 57]. Additionally, ML models are opaque and difficult to interpret. XGBs have the concept of relative importance, which measures the influence of each variable on model output [24, 58]. For example, our XGB suggests previously unreported predictors of leak including preoperative change in BMI, first assistant training level, race, ethnicity, and steroid use (Fig. 3) [7, 59]. However, unlike the LR odds ratio, relative importance does not have clear numerical or directional meaning and lacks an intuitive semantic connection to model outcomes. ANNs have no such analogous concept and are particularly difficult to interpret. In some cases, interpretable algorithms like LR may be preferable to ANN or XGB even at the expense of predictive performance.

Despite these limitations, we offer a number of innovations, particularly with respect to our ANN. It is implemented Pytorch, an industry standard framework. It makes use of a number of contemporary techniques to optimize performance and training that are common in industry but only beginning to emerge in the medical outcomes literature [23]. These include nonlinearities between layers, dropout, batch normalization, and automatic early stopping. Additionally, our ANN uses neural embeddings for categorical variables. Traditionally, categorical variables are represented as one hot for use in high-dimensional operations. By training feature vectors for each possible value of a categorical variable, we can represent values more meaningfully and in theory make better predictions [60]. This technique originated in

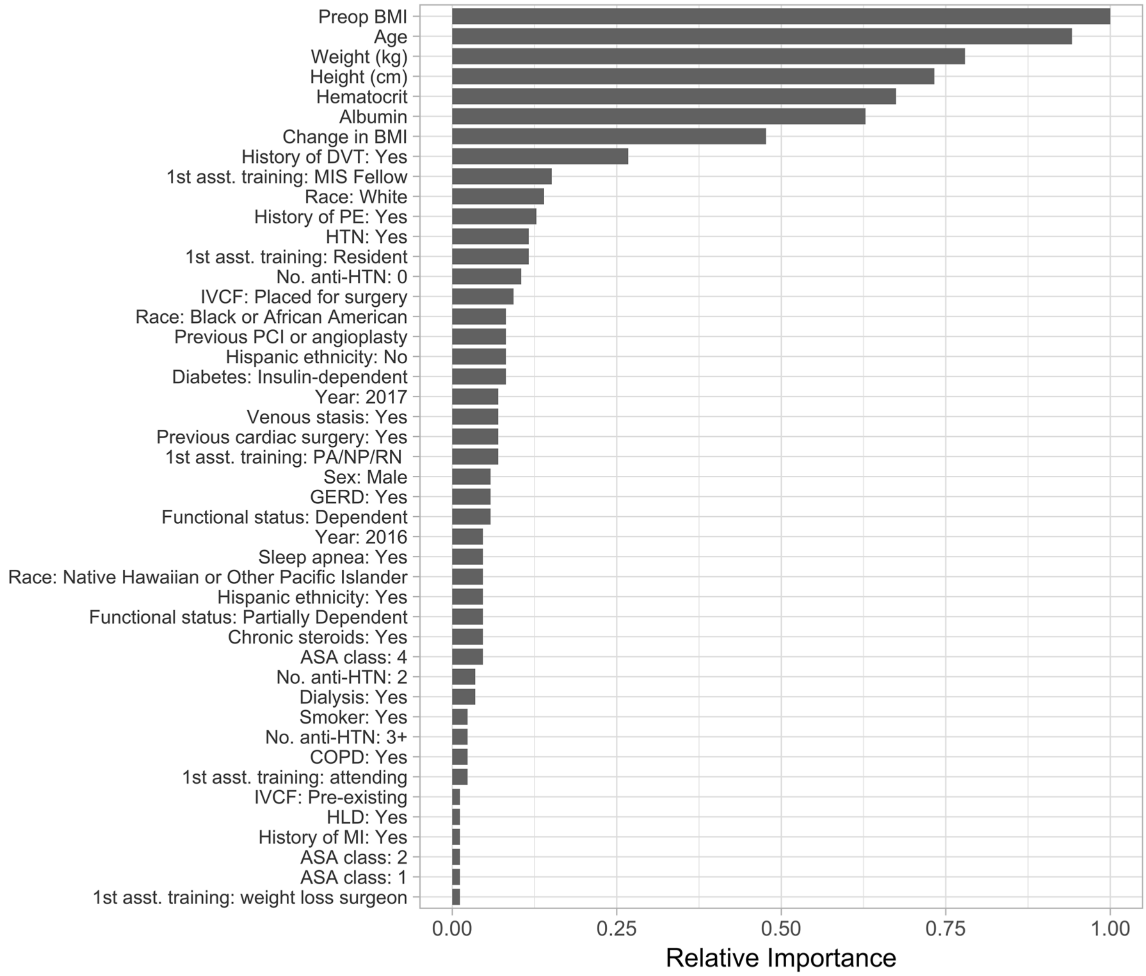

Fig. 4 Relative importance of each predictive variable in the gradient boosting machine model for predicting venous thromboembolism. *BMI* body mass index, *GERD* gastroesophageal reflux disease, *HTN* hypertension, *IVCF* inferior vena cava filter, *MIS* minimally invasive surgery, *ASA* American Society of Anesthesiology, *PE* pulmonary embolism

natural language processing [61] and has been used in commercial software [62] and data science competitions [40]. This may be its first application to surgical outcomes. Others can straightforwardly adapt our ANN to analyze any organized tabular data and modify its structure to experiment with deeper and more complicated architectures (https://github.com/jdnudel/wls_ai_open).

Artificial intelligence has the potential to transform surgery by transferring responsibility for complex cognitive and manual tasks from humans to machines, ultimately automating and amplifying the capabilities of surgical teams [20]. This study represents incremental progress toward that future and generally supports the expectation that advances in artificial intelligence and ML will meaningfully improve the performance of predictive models in surgery. To our knowledge, this is the first successful application of modern ML algorithms to characterize preoperative risk among WLS patients. Before these models can be deployed at the point of care, they must be validated in future and external cohorts. They may need to be retrained or updated with additional data in order to ensure they perform as expected in particular patient populations.

Acknowledgements The Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program (MBSAQIP) the hospitals participating in the MBSAQIP are the source of the data used herein; they have not verified and are not responsible for the statistical validity of the data analysis or the conclusions derived by the authors.

Funding Supported by the Boston University Institute for Health System Innovation and Policy and the Boston Medical Center Department of Surgery. On a separate project, Dr. Bishara was supported by a National Institute of General Medical Sciences training grant (T32 GM008440, PI: Dexter Hadley).

Compliance with ethical standards

Disclosures Drs. Nudel and Bishara are co-founders of Bezel Health, a company building software to measure and improve healthcare quality interventions. Drs. Woodson, De Geus, Srinivasan, Patil, and Hess have no conflicts of interest or financial ties to disclose.

References

1. Nguyen NT, Varela JE (2017) Bariatric surgery for obesity and metabolic disorders: state of the art. Nat Rev Gastroenterol Hepatol 14:160
2. Böckelman C, Hahl T, Victorzon M (2017) Mortality following bariatric surgery compared to other common operations in Finland during a 5-year period (2009–2013). A nationwide registry study. Obes Surg 27:2444–2451. https://doi.org/10.1007/s11695-017-2664-z
3. Fry BT, Scally CP, Thumma JR, Dimick JB (2018) Quality improvement in bariatric surgery: the impact of reducing postoperative complications on medicare payments. Ann Surg 268:22–27
4. Funk LM, Jolles S, Fischer LE, Voils CI (2015) Patient and referring practitioner characteristics associated with the likelihood of undergoing bariatric surgery: a systematic review. JAMA Surg 150:999–1005
5. Funk LM, Jolles SA, Greenberg CC, Schwarze ML, Safdar N, McVay MA, Whittle JC, Maciejewski ML, Voils CI (2016) Primary care physician decision making regarding severe obesity treatment and bariatric surgery: a qualitative study. Surg Obes Relat Dis 12:893–901
6. Vidal J, Corcelles R, Jiménez A, Flores L, Lacy AM (2017) Metabolic and bariatric surgery for obesity. Gastroenterology 152:1780–1790
7. Alizadeh RF, Li S, Inaba C, Penalosa P, Hinojosa MW, Smith BR, Stamos MJ, Nguyen NT (2018) Risk factors for gastrointestinal leak after bariatric surgery: MBASQIP analysis. J Am Coll Surg 227:135–141. https://doi.org/10.1016/j.jamcollsurg.2018.03.030
8. Mocanu V, Dang J, Ladak F, Switzer N, Birch DW, Karmali S (2019) Predictors and outcomes of leak after Roux-en-Y Gastric Bypass: an analysis of the MBSAQIP data registry. Surg Obes Relat Dis 15:396–403
9. Turrentine FE, Denlinger CE, Simpson VB, Garwood RA, Guerlain S, Agrawal A, Friel CM, LaPar DJ, Stukenborg GJ, Jones RS (2015) Morbidity, mortality, cost, and survival estimates of gastrointestinal anastomotic leaks. J Am Coll Surg 220:195–206
10. Ward-Smith P (2012) Body mass index, surgery, and risk of venous thromboembolism in middle-aged women. Urol Nurs 32:220–223
11. Klovaite J, Benn M, Nordestgaard BG (2015) Obesity as a causal risk factor for deep venous thrombosis: a M endelian randomization study. J Intern Med 277:573–584
12. ASMBS Clinical Issues Committee (2013) ASMBS updated position statement on prophylactic measures to reduce the risk of venous thromboembolism in bariatric surgery patients. Surg Obes Relat Dis 9(4):493–497. https://doi.org/10.1016/j.soard.2013.03.006
13. Aminian A, Andalib A, Khorgami Z, Cetin D, Burguera B, Bartholomew J, Brethauer SA, Schauer PR (2017) Who should get extended thromboprophylaxis after bariatric surgery. Ann Surg 265:143–150
14. Dang JT, Switzer N, Delisle M, Laffin M, Gill R, Birch DW, Karmali S (2018) Predicting venous thromboembolism following laparoscopic bariatric surgery: development of the BariClot tool using the MBSAQIP database. Surg Endosc. https://doi.org/10.1007/s00464-018-6348-0
15. Gaborit B, Aron-Wisnewsky J, Salem J-E, Bege T, Frere C (2018) Pharmacologic venous thromboprophylaxis after bariatric surgery. Ann Surg 268:e51–e52
16. Thereaux J, Lesuffleur T, Czernichow S, Basdevant A, Msika S, Nocca D, Millat B, Fagot-Campagna A (2018) To what extent does posthospital discharge chemoprophylaxis prevent venous thromboembolism after bariatric Surgery? Results from a nationwide cohort of more than 110,000 patients. Ann Surg 267:727–733
17. Kumar SB, Hamilton BC, Wood SG, Rogers SJ, Carter JT, Lin MY (2018) Is laparoscopic sleeve gastrectomy safer than laparoscopic gastric bypass? A comparison of 30-day complications using the MBSAQIP data registry. Surg Obes Relat Dis 14:264–269. https://doi.org/10.1016/j.soard.2017.12.011
18. Bahl V, Hu HM, Henke PK, Wakefield TW, Campbell DA, Caprini JA (2010) A validation study of a retrospective venous thromboembolism risk scoring method. Ann Surg 251:344–350. https://doi.org/10.1097/SLA.0b013e3181b7fca6
19. Finks JF, English WJ, Carlin AM, Krause KR, Share DA, Banerjee M, Birkmeyer JD, Birkmeyer NJ, Collaborative MBS (2012) Predicting risk for venous thromboembolism with bariatric surgery: results from the Michigan Bariatric Surgery Collaborative. Ann Surg 255:1100–1104
20. Hashimoto DA, Rosman G, Rus D, Meireles OR (2018) Artificial intelligence in surgery: promises and perils. Ann Surg 268:70–76
21. Kim JS, Merrill RK, Arvind V, Kaji D, Pasik SD, Nwachukwu CC, Vargas L, Osman NS, Oermann EK, Caridi JM (2018) Examining the ability of artificial neural networks machine learning models to accurately predict complications following posterior lumbar spine fusion. Spine 43:853–860
22. Bertsimas D, Dunn J, Velmahos GC, Kaafarani HMA (2018) Surgical risk is not linear: derivation and validation of a novel, user-friendly, and machine-learning-based predictive optimal trees in emergency surgery risk (POTTER) Calculator. Ann Surg 268:574–583. https://doi.org/10.1097/SLA.0000000000002956
23. Goto T, Camargo CA, Faridi MK, Freishtat RJ, Hasegawa K (2019) Machine learning-based prediction of clinical outcomes for children during emergency department triage. JAMA Netw Open 2:e186937. https://doi.org/10.1001/jamanetworkopen.2018.6937
24. Wong A, Young AT, Liang AS, Gonzales R, Douglas VC, Hadley D (2018) Development and validation of an electronic health record–based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. JAMA Netw Open 1:e181018. https://doi.org/10.1001/jamanetworkopen.2018.1018
25. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539
26. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp 785–794;
27. MBSAQIP. MBSAQIP participant use data file. https://www.facs.org/quality-programs/mbsaqip/participant-use. Accessed 14 Jan 2019
28. Pavlou M, Ambler G, Seaman SR, Guttmann O, Elliott P, King M, Omar RZ (2015) How to develop a more accurate risk prediction model when there are few events. BMJ 351:h3868
29. Lemaître G, Nogueira F, Aridas CK (2017) Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 18:559–563
30. Millman KJ, Aivazis M (2011) Python for scientists and engineers. Comput Sci Eng 13:9–12
31. Oliphant TE (2007) Python for scientific computing. Comput Sci Eng 9:10–20

32. Anaconda Software Distribution. Computer software. Vers. 2-2.4.0. Anaconda, Nov. 2016. Web. https://anaconda.com.
33. McKinney W (2010) Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference, vol 445. pp 51–56
34. Oliphant TE (2006) A guide to NumPy. Trelgol Publishing USA
35. Collins GS, Reitsma JB, Altman DG, Moons KGM (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 162:55. https://doi.org/10.7326/m14-0697
36. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A (2017) Automatic differentiation in Pytorch.
37. Howard JAO (2018) Fastai. GitHub. https://github.com/fastai/fastai.
38. Seth Y (2018) A neural network in PyTorch for tabular data with categorical embeddings. Let the machines learn. https://yashuseth.blog/2018/07/22/pytorch-neural-network-for-tabular-data-with-categorical-embeddings/.
39. Ng A, Katanforoosh K. CS230 Deep learning course notes and code examples
40. Guo C, Berkhahn F (2016) Entity embeddings of categorical variables. arXiv preprint. arXiv: 160406737
41. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv
42. Caruana R, Lawrence S, Giles CL (2001) Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping. Advances in neural information processing systems. MIT Press, Cambridge, pp 402–408
43. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958
44. Seabold S, Perktold J (2010) Statsmodels: Econometric and statistical modeling with python. In: Proceedings of the 9th Python in Science Conference, vol 57. p 61
45. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–845
46. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinf 12:77
47. Team RS (2015) RStudio: integrated development for R. RStudio Inc., Boston, p 42
48. Team RC (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
49. Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer, New York
50. Pollard TJ, Johnson AEW, Raffa JD, Mark RG (2018) Tableone: an open source Python package for producing summary statistics for research papers. JAMIA Open 1:26–31. https://doi.org/10.1093/jamiaopen/ooy012
51. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, Bhatt DL, Fonarow GC, Laskey WK (2017) Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. JAMA Cardiol 2:204–209. https://doi.org/10.1001/jamacardio.2016.3956
52. Johnson KW, Soto JT, Glicksberg BS, Shameer K, Miotto R, Ali M, Ashley E, Dudley JT (2018) Artificial intelligence in cardiology. J Am Coll Cardiol 71:2668–2679
53. Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, Hisamitsu T, Kojima G, Felsted J, Kakarmath S (2018) A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Med Inform Decis Mak 18:44
54. Telem DA, Yang J, Altieri M, Patterson W, Peoples B, Chen H, Talamini M, Pryor AD (2016) Rates and risk factors for unplanned emergency department utilization and hospital readmission following bariatric surgery. Ann Surg. 263:956–960. https://doi.org/10.1097/SLA.0000000000001536
55. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, Eisenmann M, Feussner H, Forestier G, Giannarou S (2017) Surgical data science for next-generation interventions. Nat Biomed Eng 1:691
56. Lee G, Rubinfeld I, Syed Z (2012) Adapting surgical models to individual hospitals using transfer learning. In: proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops, pp 57–63
57. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, Motaei A, Madkour M, Pardalos PM, Lipori G, Hogan WR, Efron PA, Moore F (2019) MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. Ann Surg 269:652–662
58. Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorobot 7:21
59. Masoomi H, Kim H, Reavis KM, Mills S, Stamos MJ, Nguyen NT (2011) Analysis of factors predictive of gastrointestinal tract leak in laparoscopic and open gastric bypass. Arch Surg 146:1048–1051. https://doi.org/10.1001/archsurg.2011.203
60. Qu Y, Cai H, Ren K, Zhang W, Yu Y, Wen Y, Wang J (2016) Product-based neural networks for user response prediction. In: Proceedings of the 2016 IEEE 16th International Conference on Data Mining. pp 1149–1154
61. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems. MIT Press, Cambridge, pp 3111–3119
62. Covington P, Adams J, Sargin E (2016) Deep neural networks for youtube recommendations. In: Proceedings of the 10th ACM conference on recommender systems. pp 191–198