Annals of Operations Research
https://doi.org/10.1007/s10479-019-03156-8

ORIGINAL RESEARCH

Predictive models for bariatric surgery risks with imbalanced medical datasets

Talayeh Razzaghi[1] · Ilya Safro[2] · Joseph Ewing[3] · Ehsan Sadrfaridpour[2] · John D. Scott[4]

Abstract

Bariatric surgery (BAR) has become a popular treatment for type 2 diabetes mellitus which is among the most critical obesity-related comorbidities. Patients who have bariatric surgery, are exposed to complications after surgery. Furthermore, the mid- to long-term complications after bariatric surgery can be deadly and increase the complexity of managing safety of these operations and healthcare costs. Current studies on BAR complications have mainly used risk scoring for identifying patients who are more likely to have complications after surgery. Though, these studies do not take into consideration the imbalanced nature of the data where the size of the class of interest (patients who have complications after surgery) is relatively small. We propose the use of imbalanced classification techniques to tackle the imbalanced bariatric surgery data: synthetic minority oversampling technique (SMOTE), random undersampling, and ensemble learning classification methods including Random Forest, Bagging, and AdaBoost. Moreover, we improve classification performance through using Chi-squared, Information Gain, and Correlation-based feature selection techniques. We study the Premier Healthcare Database with focus on the most-frequent complications including Diabetes, Angina, Heart Failure, and Stroke. Our results show that the ensemble learning-based classification techniques using any feature selection method mentioned above are the best approach for handling the imbalanced nature of the bariatric surgical outcome data. In our evaluation, we find a slight preference toward using SMOTE method compared to the random undersampling method. These results demonstrate the potential of machine-learning tools as clinical decision support in identifying risks/outcomes associated with bariatric surgery and their effectiveness in reducing the surgery complications as well as improving patient care.

Keywords Imbalanced data · Risk prediction · Clinical decision support · Bariatric surgery

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10479-019-03156-8) contains supplementary material, which is available to authorized users.

✉ Talayeh Razzaghi
talayehr@nmsu.edu

Extended author information available on the last page of the article

Published online: 01 February 2019

1 Introduction

Being known as the blood sugar level for a prolonged period, Diabetes Mellitus (or just Diabetes in a shorter term) is now growing at an Epidemic rate in the United States according to American Diabetes Association (2015). Studies show that Diabetes Mellitus is among the leading causes of disability, morbidity, and mortality in the United States (Almdal et al. 2004; Centers for Disease Control and Prevention 2011; Kannel and McGee 1979; Stamler et al. 1993). In a broader scale, the World Health Organization estimated that about 422 million adults were living with Diabetes in 2014 (World Health Organization 2016). Although the occurrence rate of Diabetes-related complications has significantly been reduced due to recent endeavors in glycemic control and cardiovascular risk factor management, the rise of Diabetes prevalence has solely lead to growing numbers of macrovascular and microvascular disease incidents over the last few years. Diabetes appears in three forms as stated in American Diabetes Association (2006): Diabetes Type 1, which results from insulin deficiency and accounts for 5–10% of diabetic patients, Diabetes Type 2, which is previously referred as non-insulin dependent diabetic patients and accounts for 90–95% of diabetic diagnoses and Gestational diabetes.

Patients with Diabetes Mellitus Type 2 (T2DM) often suffer from Obesity-related illnesses. As a closely-related metabolic syndrome, Obesity is also associated with several health risks. According to a recent report published by Center for Disease Control and Prevention, more than one-third of adults in the United States suffer from Obesity (Ogden et al. 2015). According to a study by Cawley and Meyerhoefer (2012), the national medical care costs of Obesity-related illnesses in adult pass more than \$200 billion a year. Obesity, which is defined as having a Body Mass Index (BMI) of $30\,\text{kg/m}^2$, has been shown to be a serious health risk factor for both T2DM and Cardiovascular diseases. For example, the results from a clinical study conducted by Daousi et al. (2006) show that Obese patients with T2DM suffer from worse cardiovascular risk factors compared to other diabetic patients without obesity. However, a limited number of works have studied the impact of Obesity on arising cardiovascular consequences on diabetic patients (Johnson et al. 2013, 2012).

Bariatric surgery (BAR) has been shown to be a successful therapy for T2DM patients with Obesity, which can lead to significant and persistent weight loss (Grundy et al. 1991; Brolin 1996; Buchwald 2005). According to some studies, BAR has proved to result in complete remission of T2DM in about 75–80% of patients (Buchwald et al. 2004, 2009). BAR is performed through one of these four distinct procedures: Rouxen-Y gastric bypass (RYGB), Gastric banding (LAGB), biliopancreatic diversion (BPD), and sleeve gastrectomy (SG).

However, limited studies (Pories 2008; DeMaria et al. 2007) have addressed the mid- to long-term outcomes/risks of diabetes and obese patients after bariatric surgery. Risk scoring using logistic regression analysis is the most commonly used technique in bariatric surgery risk studies. DeMaria et al. (2007) developed a risk scoring system by logistic regression to identify the most important predictors of increased rate of mortality after surgery.

Since the majority group (patients who have complications after surgery) dominates the behavior of logistic regression analysis, it might not be an appropriate method for imbalanced classification problems (King and Zeng 2001). Failing to correctly identify patients who are at risk of complications after surgery can lead to significant costs and even loss of life. Hence, it is necessary to develop classification models that yield accurate detection of complication/risk events. Because such models will benefit clinicians to improve patient outcomes after surgery and provide cost-effective care for high-risk patients.

Our work lies in the medical pattern recognition framework, which is known to be highly imbalanced, i.e., the instances of interest in the dataset are relatively rare. Examples are Intensive Care Unit (ICU) infection detection events (Roumani et al. 2013, 2018), medical diagnoses (Khalilia et al. 2011; Alexe et al. 2003; Şeref et al. 2017), adverse drug events (Taft et al. 2009; Sarker and Gonzalez 2015), bleeding detection in endoscopic video (Deeba et al. 2016), and so on. Bariatric surgery results lie in this group as well because BAR risks/complications are very skewed and the high-risk groups often form the minority class. However, to the best of authors' knowledge, there is no comprehensive study that handles the imbalanced nature of the bariatric surgical risk prediction problems.

In this work, we study the merit of using imbalanced classification techniques to predict the outcomes appearing in T2DM patients with Obesity, who have undergone BAR. In particular, we consider Stroke, Diabetes, Angina, Blindness, Myocardial Infarction, Heart Failure, and Death as the most-potential outcomes of BAR (Johnson et al. 2013) and construct predictive models by solving classification problems for each of them.

This paper is organized as follows. In Sect. 2, we describe the approaches and methods, measures, and data used in the study. In Sect. 3, we provide empirical results and discussion. Finally, in Sect. 4 we give our conclusions, and ideas for future research.

2 Materials and methods

The Premier Healthcare Database is one of the largest U.S. healthcare datasets, which gathers the data from 700+ hospitals across U.S. and contains clinical and health-economic data. This database includes both inpatient and outpatient visits and records the costs, diagnoses, and procedures associated with each visit as well as the demographic information about the patients.

Figure 1 shows the overview of our method. In the following subsections, we discuss the details of our method.

2.1 Data preparation

In this section, we discuss the overview of our data preparation step. (We present the detailed results in Sect. 4.) For this research work, we limit our study to the T2DM patients with Obesity, who have undergone BAR. We exclude patients who had no ICD-9-CM diagnosis code associated with moderate or severe obesity or contained missing data. Using the Premier Healthcare Database, three categories of data are extracted for each of these patients. First, we select a number of patient-specific attributes including age, race, gender, ethnicity, the insurance provider, and the marital status for each such patient. Second, we consider an array of candid health-related attributes that reflect the patient's clinical history. As stated in Sect. 2.2., these candid features will be analyzed via feature selection methods that pick the most influential set of features for the related classification problem. Third, we extract the information about seven specific outcomes as the most potential outcomes of BAR [as stated by Johnson et al. (2013)]. For this purpose, only outcomes that occur after BAR date are identified. It is worth mentioning that we do not include missing or incompatible data in our study.

Remark 1 We need to consider remedies for the variation in each patient's age and marital status considering the fact that the data is related to a period of 4 years. These remedies are stated in Sect. 3 of this paper.

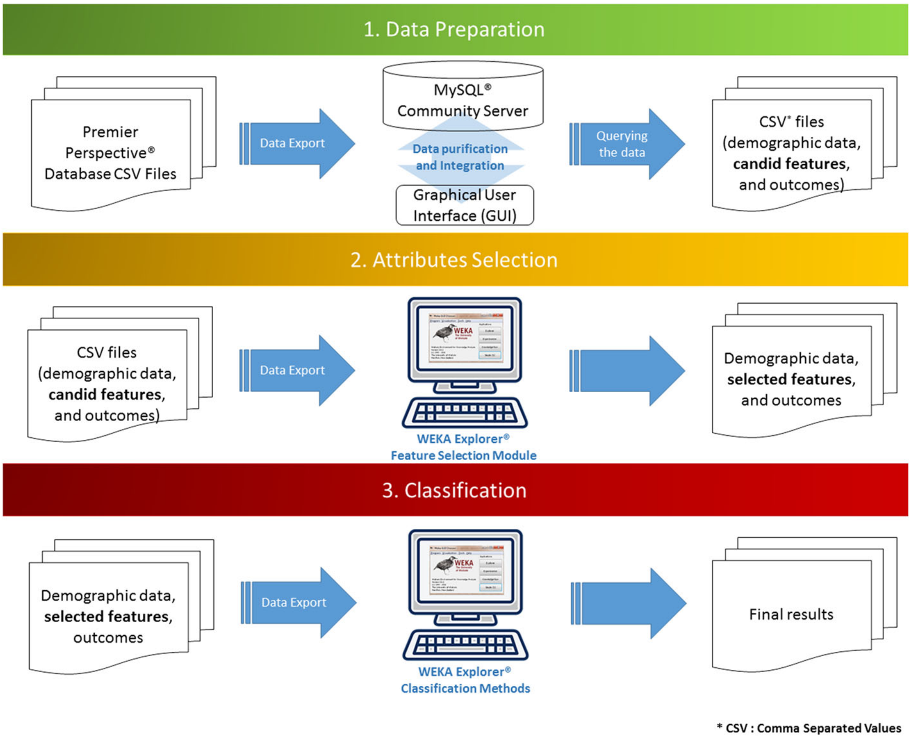
1. Data Preparation
Premier Perspective® Database CSV Files
Data Export
MySQL® Community Server
Data purification and Integration
Graphical User Interface (GUI)
Querying the data
CSV* files (demographic data, candid features, and outcomes)
2. Attributes Selection
CSV files (demographic data, candid features, and outcomes)
Data Export
WEKA Explorer® Feature Selection Module
Demographic data, selected features, and outcomes
3. Classification
Demographic data, selected features, outcomes
Data Export
WEKA Explorer® Classification Methods
Final results
* CSV : Comma Separated Values

Fig. 1 The overview of our approach

2.2 Feature selection

Given each outcome, not all features are of significant influence on the outcome. In fact, considering irrelevant features may produce less-accurate classification results and can lead to biased predictive models. In addition, such features may result in overfitting, which might have negative impacts on the accuracy of a model. Hence, feature selection is recommended prior implementing any data-mining algorithm. A new optimization-based framework, called Support Feature Machine (SFM), has been found efficient for feature selection in medical data classification (Fan and Chaovalitwongse 2010). Depending on how the feature selection search is combined with the classification model, feature selection techniques can be categorized into three strategies: filter techniques, wrapper techniques, and embedded techniques (Saeys et al. 2007).

In filter techniques, each feature's relevance score is computed based on the inherent properties of the data. The most relevant features are then selected for the next step, and the features at the bottom of the scoring list are eliminated. The most common filter selection algorithms are Information Gain (IG) (Inza et al. 2000), Chi-square (Zheng et al. 2015), and Correlation-based Feature Selection (CFS) (Zheng et al. 2015; Hall 2000).

Wrapper methods act based on the appropriateness of subsets of the features (unlike the filter methods that compute the advantage (i.e., the relevance score) for each feature). These techniques first determine the space of feature subsets followed by construction of various combinations of features (stored as subsets). Upon performing the training step of a specific classification algorithm (e.g., Naïve Bayes, bagging, etc.), one can compute the most relevant subset of features. Hence, it is said that we “wrap” a search method around a specific classification model to examine the entire space of the feature subsets. Note that as the number of features increases, the space of feature subsets exponentially grows, which can significantly affect the performance of the wrapping-based methods. Hence, heuristics become more appropriate choices to tackle real-world problems. The wrapper methods commonly use randomized search heuristics (Blanco et al. 2004; Jirapech-Umpai and Aitken 2005; Li et al. 2001; Ooi and Tan 2003) and sequential search techniques (Inza et al. 2004; Xiong et al. 2001).

In embedded techniques, the classifier construction step also involves a search method for an optimal subset of features within the combined space of the feature subsets and hypotheses. Similar to wrapper methods, these methods are also specific to a given learning algorithm, although they are less computationally expensive (Saeys et al. 2007).

For this work, we choose filter methods for the feature selection step. This is justified by the fact that filter techniques are known to be computationally fast and, hence, appropriate for real-world datasets. Moreover, they are independent of the choice of the classification techniques, which can significantly reduce the computations required for the feature selection step. Yang and Pedersen (1997) stated that IG and Chi-square performed successfully in the multi-class classification framework.

2.2.1 Information Gain

The Information Gain (IG) algorithm measures the reduction in entropy when the feature is present. The concept of entropy is used as a measure of the uncertainty of a random variable. The entropy of a variable X is calculated as

$$H(X) = -\sum_{i} P(x_i)\log_2(P(x_i)) \tag{1}$$

where the prior probability for the value of X is denoted by $P(x_i)$. After observing values of another variable Y, the entropy of X is given by,

$$H(X|Y) = -\sum_{j} P(y_j)\sum_{i} P(x_i|y_j)\log_2(P(x_i|y_j)) \tag{2}$$

where the $P(x_i|y_i)$ is the posterior probability of $X = x_i$ given the data $Y = y_i$. Further information about X provided by Y is measured by the decrease of entropy of X and thus is defined as *information gain* (IG) (Quinlan 2014):

$$IG(X|Y) = H(X) - H(X|Y) \tag{3}$$

Thus, if $IG(X|Y) > IG(Z|Y)$, it denotes that the feature Y is more correlated to the feature X compared to feature Z.

2.2.2 Chi-square

Chi-square (χ^2), another popular feature selection method, is used to select relevant features by considering the classes. In this approach, the continuous-valued features are discretized into several intervals. Assume that N is the total number of examples and N_{ij} is the number of examples belongs to the class C_i and the jth interval. M_{lj} is the number of examples in the jth interval, and l is the number of intervals. The expected frequency of N_{ij} is given by,

$$E_{ij} = \frac{M_{lj}|C_i|}{N} \tag{4}$$

The χ^2 statistic of a feature is defined as,

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{l} \frac{(N_{ij} - E_{ij})^2}{E_{ij}} \tag{5}$$

The larger value of χ^2 reflects that feature is more informative.

2.2.3 Correlation-based feature selection (CFS)

CFS selects the best feature subset with respect to the predictive performance of individual feature as well as the amount of redundancy among them. The correlation between a subset of features and classes and the inter-correlation between the features are calculated by correlation coefficients. As the correlation between features and classes increases and the inter-correlation decreases, the relevance of a subset of features increases (Hall 1999). CFS typically applies search methods such as forward selection, best-first search, bi-directional search, backward elimination, and genetic search. The merit of a feature subset (S) with k features is given by

$$Merit_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{6}$$

where k is the number of features, $\overline{r_{cf}}$ is the average of the correlations between the subset features and the class label, and $\overline{r_{ff}}$ is the average inter-correlation between subset features (Hall 1999).

2.3 Classification approaches and methods

The construction of a classification model for imbalanced bariatric surgical data is performed upon finishing the data preparation and feature selections steps. The presence of imbalanced classes often results in serious bias in the performance metrics. In fact, most of the state-of-the-art data-mining techniques tend to obtain a decision boundary that is biased toward the majority class. As a result, if imbalanced-ness is neglected, the technique tends to misclassify instances from the minority class (e.g., the rare outcomes of BAR), while it is highly crucial to identify the minority class instances.

2.3.1 Approaches

To cope with the imbalanced-ness issue, several remedies have been suggested including undersampling (Batista et al. 2004), oversampling (Chawla et al. 2002), the cost sensitive algorithms (Zadrozny et al. 2003) and ensemble learning methods (Polikar 2006). Both undersampling and oversampling methods try to balance the two classes through either decreasing the size of the majority class or increasing the size of the minority class. Cost-sensitive learning methods employ larger penalty for misclassification of minority class (compared to the one of majority class). This prohibits generation of boundaries biased to the majority class. Although being very precise, the main difficulty in using these methods arise in the computation of appropriate penalty values. The main idea of ensemble-based classifiers is to aggregate the predictions obtained by applying several base classifiers into an imbalanced data set with the hope of getting improved results compared to each classifier's result (Rokach 2010). Adaptive Boosting (AdaBoost) (Schapire 1990), Bagging (Breiman 1996), and Random Forest (Liaw and Wiener 2002) are among the mostly-used algorithms in the ensemble learning framework.

Random undersampling (RUS) removes the instances from the majority class randomly until the desired majority to minority class ratio is reached.

Synthetic minority over-sampling technique (SMOTE) generates a synthetic instance by interpolating k instances (for a given integer value k) of the minority class that lies close enough to each other (López et al. 2013). Oversampling methods aim to achieve the desired ratio by creating "synthetic" instances of the minority class.

Adaptive Boosting (AdaBoost) Freund and Schapire (1995) is the most well-known algorithm in the boosting family (Schapire 1990). AdaBoost trains each classifier sequentially using the entire dataset. After each iteration, it concentrates more on problematic observations that were misclassified in the previous iteration. It aims to classify these observations correctly in current iteration through a weighting strategy. All observations get equal weights in the first round of training, then at each iteration, AdaBoost increases the weights of incorrectly classified examples while decreases the weights of correctly classified examples. Moreover, this algorithm assigns another weight to each classifier based on its overall accuracy. Better classifiers receive higher weights. Then the class label of a new example is determined by selecting majority of weighted votes that are given by each classifier.

Bagging or the bootstrap aggregating to construct ensembles was first introduced by Breiman (1996). It uses bootstrapped replicas of the initial training set to train different classifiers. Finally, when an unknown example is given to each classifier, the class label is identified by a majority or weighted vote. Algorithm 1 demonstrates the pseudocode for Bagging.

Algorithm 1 Bagging

```
Input: $S$: Training set, $N$: Bootstrap size, $T$: Number of iterations, $I$: weak classifier
for $k = 1 : T$ do
    $S_k \leftarrow RandomSampleReplacement(N, S)$
    $h_k \leftarrow I(S_k)$
end for
Output: An ensemble by the Majority voting scheme, $H(x) = sign(\sum_{k=1}^{T} h_k(x))$ where $h_k \in \{-1, 1\}$
    are base classifiers.
```

Random Forest (RF) Breiman (2001) is an ensemble learning method that builds a set of decision tree classifiers to find the label of a new example by voting for the most popular class. For each decision tree classifier, Bagging is used on the original training data to create many copies of it. Each decision tree classifier differs from the rest in a way that the split on each node is based on the best feature chosen from a randomly selected set of all candidate features. Finally, the class label of a new instance is assigned through majority voting among all votes (i.e., predicted label) given by each tree (small base classifier) of an RF. Algorithm 2 shows the pseudocode for RF.

Algorithm 2 Random Forest

```
Input: $S$: Training set, $F$: Feature Set, $T$: Number of trees in forest
function RandomForest($S$, $F$)
    for $i = 1 : T$ do
        $S_i \leftarrow bootstrapSample(S)$ (select a bootstrap sample from S)
        At each node:
            $f \leftarrow$ randomly select a subset of the features from F
            Split on best feature in f
            return $h_i$
    end for
end function
Output: An ensemble by the Majority voting scheme, $H(x) = sign(\sum_{i=1}^{T} h_i(x))$ where $h_i \in \{-1, 1\}$
    are tree classifiers.
```

2.3.2 Methods

In this paper, we employ six of the most popular classification methods coupled with (under/over) sampling methods as remedies to treat the imbalanced nature of the data. These methods are (1) Naïve Bayes, (2) Radial Basis Function Neural Network (RBFNN), (3) 5-Nearest Neighbors (5NN), (4) Decision Trees (C4.5 Algorithm also known as J48 Algorithm), (5) Support Vector Machines (SVMs), and (6) Logistic Regression (LR). The reader is referred to the book authored by Friedman et al. (2001) to obtain more information about the aforementioned techniques. In addition, we employ a hybrid approach. Our motivation originates from studies that advocate combining the (under/over) sampling procedures with the ensemble learning algorithms (Galar et al. 2012, 2013). In particular, we study six such approaches, which are obtained by combining each of (under/over) sampling methods with ensemble learning techniques including Random Forest (RF), Bagging, and AdaBoost classifiers. Our initial experimental studies (refer to A5) demonstrated the superiority of using Radial Basis Function (RBF) kernel over linear kernel when implementing SVM algorithm. So, in our implementations, we only work with SVMs equipped with RBF kernels. The nonlinear kernel is often prohibitive on too big data because of the complexity. In such cases, acceleration techniques such as multilevel SVM (Razzaghi et al. 2016; Razzaghi and Safro 2015) can be used.

3 Results and discussion

3.1 Classifier evaluation metrics

Several metrics have been proposed to validate the results of a classification algorithm. Accuracy, Precision, Sensitivity, Specificity, G-mean, F-Measure, and the area under the Receiver Operating Characteristic (ROC) curve are few of common metrics, which are mainly computed from the Confusion Matrix as depicted in Table 1 (Gu et al. 2009).

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad \text{Specificity} = \frac{TN}{TN + FP} \tag{8}$$

$$\text{G-mean} = \sqrt{Sensitivity * Specificity} \tag{9}$$

$$\text{F-measure} = \frac{2TP}{2TP + FP + FN}, \quad \text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

Table 1 Confusion matrix for binary classification problem

	Positive class	Negative class
Positive class	True positive (TP)	False positive (FP)
Negative class	False negative (FN)	True negative (TN)

Note that since the bariatric surgical complication datasets are highly imbalanced, some of the metrics above may produce misleading interpretations [e.g., "Accuracy" performance metric as stated by Galar et al. (2012)]. Here, we put more emphasis on the G-mean performance metrics due to the fact that it reflects both "Specificity" and "Sensitivity" measures. Moreover, we report the area under ROC curve, which plots "Sensitivity" versus "Specificity." The ROC area measures the ability of the classifier to classify the majority and minority classes correctly.

3.2 Results

In this section, we present the detailed results of our study including the data preparation, feature selection, and classification. We start by providing details about the data preparation regarding the patients' characteristics (including both demographic and clinical attributes) and discuss the outcomes next. Finally, we state the results of our implementations.

3.2.1 Patients characteristics

The Premier Healthcare Database contains no patient-identifiable information, and the patients cannot be tracked across institutions; their visits to the same hospital can be tracked using the *MEDRECKEY*; the *PATKEY* would represent each individual visit to that institution and would be different for each visit. Through a careful study of the database records between 2011 and 2014, we have observed more than 4M patients' visits along with more than 50M records about the diagnoses/procedures occurred during those visits. To store and query such a massive amount of data, we employ MySQL@ Community Server (as depicted in Fig. 1). By employing the ICD-9 codes for T2DM, Obesity, and BAR, we also extract only those records that belong to T2DM patients with Obesity, who have undergone BAR. (Please refer to Table A1 in the Appendix for the ICD-9 codes.) These include 11636 patients. In the rest of this section, we limit our focus to such records.

In selecting the most significant features for our predictive model, we heavily rely on previous studies in Johnson et al. (2012, 2013) and Stamler et al. (1993), which are conducted by medical researchers. For the demographic data, we collect each patient's gender, ethnicity, insurance provider, age, and marital status. That is, five candid attributes of demographic data has been stored for each patient. Note that these attributes are recorded at all visits that each patient makes. Therefore, as mentioned in Remark 1, variations in the age and marital status can be substantial and must be taken into account. We explain our remedies to cope with this issue next. Within the records of the patients under of our study, we observed a variation of less than 2 for the age, so we decided to work with the average of the age feature. Regarding the marital status, three possible states had originally been defined within the $MartStat$ field in the Premier Healthcare Database: (1) Married, (2) Single, and (3) Other. However, we observed that a patient's marital status might change between the date of BAR and the outcome under study. Hence, we considered six more possible values for the marital status, which reflect the possible change in the marital status. For each patient, we first let the $MartStat$ field be the marital status at the time of bariatric surgery. Depending

Table 2 Patient demographic attributes

Feature	Value	Frequency	Feature	Value	Frequency
Gender	Female	8259 (70.9%)		Married (M)	3867 (33.2%)
	Male	3377 (29.1%)		Single (S)	5834 (50.1%)
Ethnicity	White	7780 (66.8%)		Other (O)	1839 (15.9%)
	Black	1553 (13.4%)	Marital status	M to S	15 (0.1%)
	Other	2303 (19.8%)		M to O	21 (0.2%)
Insurance	Medicare	3305 (28.4%)		S to M	27 (0.2%)
	Medicaid	1130 (9.7%)		S to O	13(0.1%)
	Managed care	5208 (44.8%)		O to M	8 (0.1%)
	Commercial	1101 (9.4%)		O to S	12 (0.1%)
	Self-pay	221 (1.9%)	Age	Varies in 13–86	
	Other	671 (5.8%)		Years old.	

Table 3 Patient clinical attributes

Feature	Frequency	Feature	Frequency
COPD	2978 (25.6%)	Coronary artery disease	1179 (10.1%)
Diabetic manifestations	416 (3.6%)	Transient ischemic attack	11 (0.1%)
Tobacco abuse	590 (5.1%)	Sleep apnea	2337 (20.1%)
Hypertension	9074 (77.9%)	Dyslipidemia	7457 (64.9%)

on the outcome under study, if the patient maintains the same marital status at the date of the outcome occurred, we leave this field unchanged. Otherwise, depending on the change in the marital status, we let $MartStat$ take one of the following six values: (4) Married to Single, (5) Married to Other, (6) Single to Married, (7) Single to Other, (8) Other to Single, and (9) Other to Married. To include the clinical history of the patients into our analysis, we also collect information about eight comorbid conditions/diseases [as stated in Johnson et al. (2012)] as candid features using their associated ICD-9 codes. We include these candid clinical-based features provided that the condition/diagnosis have occurred earlier than the BAR date. Tables 2 and 3 show the general information regarding the patients' demographic data and clinical attributes, respectively.

It is worth mentioning that for this period of study, the patients under study were between 13 and 86 years old and had no previous history of myocardial infarction (MI), angina, congestive heart failure, stroke, and blindness in at least one eye.

3.2.2 Outcomes

Based on the study by Johnson et al. (2013), we refer to seven common outcomes that can occur after any of the four standard BAR procedures (as mentioned in Sect. 2). Table 4 state these outcomes along with their frequencies within the patients under study. According to Table 4, the number of patients in the class of risks/outcomes (positive class) is extremely fewer than the number of patients of the class of no risks/outcomes (negative class). Hence, this data is highly imbalanced and justifies our choice of using special classification methods. Note that none of the "Blindness," "Myocardial Infarction," and the "Death" outcomes yield

Table 4 Outcomes of BAR and their frequency among patients under study

Label	Frequency
Diabetes	1543 (13.2%)
Heart failure	396 (3.4%)
Stroke	43 (0.3%)
Angina	51 (4.4%)
Myocardial infarction	4 (0.03%)
Death	0
Blindness	6 (0.05%)

a reasonable-size data set for data-mining techniques. Hence, we limit our study to four outcomes: Diabetes, Angina, Heart Failure, and Stroke.

3.2.3 Implementation

In this section, we describe experimental results using both random undersampling and oversampling (SMOTE) in combination with base classifiers and widely-used ensemble learning algorithms (as stated in Sect. 2.3). We implement our approaches using Waikato Environment for Knowledge Analysis known as WEKA (Witten et al. 2016). WEKA is a free license workbench developed to perform predictive modeling and data analysis and includes several modules. In particular, we use WEKA Explorer module to implement various classification algorithms. We also employ feature selection tool available in WEKA.

For both undersampling and oversampling methods, the desired ratio of the class sizes is considered to be 50:50. For the oversampling (SMOTE) technique, we employ 5-nearest neighbors (5NN) algorithm to create new instances of the minority class. Euclidean distances are used to compute the necessary closeness values for the 5NN technique embedded within SMOTE method. Finally, we employ 10-fold cross-validation to calculate the estimates of the performance metrics.

Our work lies in the context of the one-against-all multi-class classification problem. The classification task is conducted to find out whether a patient suffers from an outcome or not (no matter if the patient develops other issues or not). Table 5 reports the performance metrics for the three feature selection methods when applied to the classification problems for the Diabetes outcome. Note that for each classification algorithm and each performance metric, we report the results of both undersampling (in column "U") and oversampling (in column "O") methods. We have reported similar information regarding Angina, Heart Failure and the other three outcomes in the Appendix in tables A2, A3, and A4, respectively.

To obtain an idea about the performance of each feature selection method, we also report the selected number of features. In particular, for each outcome, we report the number of original features as well as the number of selected features that have been employed by the best classifier (which is determined as having the highest value of G-mean performance metric for that outcome). Table 6 demonstrate such information.

3.3 Discussion

According to Tables 5, A2–A4, the oversampling method dominates the undersampling method when both considered for ensemble learning classifiers and the same feature selection

Table 5 Results for diabetes outcome

Classifier	Feature selection	Acc		G-mean		Precision		Sensitivity		Specificity		F-measure		ROC area	
		U	O	U	O	U	O	U	O	U	O	U	O	U	O
Naïve Bayes	Chi-squared	0.61	0.61	0.60	0.60	0.59	0.59	0.69	0.71	0.53	0.51	0.64	0.65	0.64	0.65
	Information Gain	0.61	0.61	0.60	0.60	0.59	0.59	0.69	0.71	0.53	0.51	0.64	0.65	0.64	0.65
	CFS	0.61	0.59	0.61	0.58	0.60	0.61	0.69	0.50	0.53	0.68	0.64	0.55	0.64	0.64
RBFN	Chi-squared	0.56	0.57	0.56	0.52	0.56	0.55	0.57	0.81	0.56	0.34	0.57	0.66	0.59	0.61
	Information Gain	0.56	0.58	0.56	0.54	0.56	0.56	0.57	0.79	0.56	0.38	0.57	0.65	0.59	0.61
	CFS	0.57	0.55	0.56	0.55	0.56	0.56	0.69	0.54	0.45	0.57	0.61	0.55	0.59	0.59
5NN	Chi-squared	0.57	0.74	0.56	0.74	0.56	0.73	0.69	0.77	0.46	0.71	0.62	0.75	0.61	0.82
	Information Gain	0.57	0.74	0.56	0.74	0.56	0.72	0.69	0.77	0.46	0.70	0.62	0.75	0.61	0.82
	CFS	0.60	0.72	0.59	0.70	0.59	0.66	0.71	0.90	0.50	0.54	0.64	0.76	0.63	0.81
C4.5 (J48)	Chi-squared	0.60	0.78	0.59	0.78	0.58	0.78	0.68	0.78	0.51	0.78	0.63	0.78	0.61	0.78
	Information Gain	0.60	0.77	0.59	0.77	0.58	0.77	0.68	0.78	0.51	0.77	0.63	0.77	0.61	0.84
	CFS	0.60	0.73	0.60	0.72	0.59	0.69	0.66	0.84	0.54	0.62	0.63	0.76	0.61	0.81
SVM	Chi-squared	0.61	0.70	0.60	0.70	0.59	0.70	0.69	0.72	0.53	0.68	0.64	0.71	0.61	0.70

Table 5 continued

Classifier	Feature selection	Acc		G-mean		Precision		Sensitivity		Specificity		F-measure		ROC area	
		U	O	U	O	U	O	U	O	U	O	U	O	U	O
	Information Gain	0.61	0.71	0.60	0.71	0.59	0.7	0.69	0.73	0.53	0.69	0.64	0.71	0.61	0.71
	CFS	0.60	0.70	0.60	0.69	0.59	0.66	0.68	0.81	0.53	0.59	0.63	0.73	0.60	0.70
LR	Chi-squared	0.61	0.62	0.61	0.62	0.6	0.61	0.64	0.64	0.58	0.59	0.62	0.63	0.65	0.67
	Information Gain	0.61	0.62	0.61	0.62	0.6	0.61	0.64	0.64	0.58	0.59	0.62	0.63	0.65	0.66
	CFS	0.60	0.59	0.60	0.59	0.6	0.59	0.63	0.58	0.58	0.60	0.61	0.59	0.65	0.63
Random Forest	Chi-squared	0.56	0.83	0.56	0.83	0.56	0.80	0.64	0.87	0.49	0.79	0.59	0.83	0.60	0.90
	Information Gain	0.56	0.83	0.56	0.82	0.56	0.80	0.64	0.86	0.49	0.79	0.59	0.83	0.60	0.90
	CFS	0.60	0.83	0.59	0.82	0.59	0.78	0.69	0.92	0.51	0.74	0.63	0.84	0.63	0.88
AdaBoostM1	Chi-squared	0.59	0.66	0.59	0.65	0.59	0.63	0.60	0.75	0.59	0.56	0.59	0.69	0.63	0.71
	Information Gain	0.59	0.66	0.59	0.65	0.59	0.63	0.60	0.75	0.59	0.56	0.59	0.69	0.63	0.71
	CFS	0.59	0.63	0.59	0.63	0.59	0.66	0.60	0.54	0.59	0.73	0.59	0.60	0.63	0.70
Bagging	Chi-squared	0.59	0.84	0.58	0.84	0.57	0.81	0.67	0.90	0.50	0.78	0.62	0.85	0.62	0.91
	Information Gain	0.59	0.84	0.58	0.84	0.57	0.81	0.67	0.90	0.50	0.78	0.62	0.85	0.62	0.91
	CFS	0.61	0.84	0.60	0.84	0.59	0.79	0.69	0.94	0.53	0.74	0.64	0.86	0.63	0.91

Table 6 Summary of the result for best feature selection and classification of each outcome

Outcome	Classifiers	Feature selection	$\|Selected\ Features\|$	G-mean	ROC area
Diabetes	Bagging	Chi-squared	24	0.84	0.91
		IG	21	0.84	0.91
		CFS	8	0.84	0.91
Angina	RF	Chi-squared	25	1.00	1.00
		IG	21	1.00	1.00
		CFS	14	1.00	1.00
	Bagging	Chi-squared	25	1.00	1.00
		IG	21	1.00	1.00
Heart failure	RF	Chi-squared	22	0.95	0.98
		IG	20	0.95	0.98
Stroke	RF	Chi-squared	25	1.00	1.00
		IG	21	1.00	1.00
		CFS	13	1.00	1.00

The number of original features is 25

technique. We observe, for example, the former method can produce an improvement of about 30% in some cases. For example, for the "Heart Failure" outcome reported in Table A3, compare the methods using the G-mean when they are used within Random Forest classifier with any feature selection technique. The oversampling method also outperforms the undersampling method when both considered for 5NN, C4.5, and SVM base classifiers with any feature selection technique. We relate this result to the loss of valuable information that is more likely in undersampling technique due to removing instances from the majority class (which could negatively affect building an accurate model). For the RBFN method, however, we observe slightly better G-mean values for the undersampling method. The difference between methods when employed with Naïve Bayes and LR are indistinguishable (with the Stroke outcome as the exception for Naïve Bayes). The "CFS" feature selection method is very slightly outperformed by the "Chi-Squared" and the "Information Gain" feature selection methods in some cases (i.e., combinations of classifiers and sampling method) for all outcomes, although they remain indistinguishable. An interesting exception occurs when this behavior is studied for RBFN base classifier, which reveals the superiority of "CFS" to the other two feature selection methods for all outcomes. This agrees with some present studies (Karegowda et al. 2010).

In general, the ensemble learning classifiers yield better performance metrics compared to base classifiers when studying all outcomes. The highest performance values for these classifiers are attained when oversampling method is employed. Within ensemble learning classifiers, we observe that "AdaBoost" classifier is almost always outperformed by the "Random Forest" and "Bagging" classifiers. This result is held for all performance metrics. Within base classifiers, the "5NN" and the "C4.5" classifiers result in best G-mean values in all outcomes followed by the "SVM" classifier. We note that while the best performance of the RBFN classifier may occur using any of the sampling methods, it is always outperformed by the "Naïve Bayes" and "LR" base classifiers. The difference between the two latter classifiers, however, is not substantial in terms of the G-mean performance metric.

Based on Tables 5, A2–A4, the best approach to classify the Diabetes outcome is the combining of any of feature selection methods and Bagging classifier, which produces 84%

classification G-mean and 91% ROC area value. In the classification of Angina outcome, both Random Forest classifier (independent from our choice of feature selection method) and Bagging classifier (when combined with either Information Gain or Chi-squared feature selection methods) produce the highest classification G-mean and ROC area values. The best methods for the classification of Heart Failure outcome data set are combining either Information Gain or Chi-squared feature selection methods with Random Forest classifier, which yields 95% G-mean and 98% ROC area values. The next best choice here is obtained by combining the Information Gain or Chi-squared feature selection with Bagging classifier. The Random Forest also yields the highest performance values for the classification of Stroke independent from our choice of feature selection method. Interestingly, both the Bagging and the C4.5 classifiers are ranked the second best alternative, in this case, yielding 99% G-mean value and 100% for the ROC area value. Table 6 reports the best approaches for each outcome. It also states the actual number of selected features that have been used by the feature selection methods.

4 Conclusion

This paper proposes the application of imbalanced classification techniques to identify bariatric surgery's complications for each patient. By extracting the required data sets from the Premier Healthcare Database, we investigate various data-mining methods to determine the risk group of a particular patient, including commonly-used base classifiers as well as ensemble learning and sampling methods to mitigate the effects of the imbalanced data set. Furthermore, we compare the advantage of using well-known feature selection methods prior to classification. Our results show that the combination of a suitable feature selection method with ensemble learning methods equipped with Oversampling (SMOTE) method can achieve higher performance metrics.

Acknowledgements This work was funded in part by Clemson University—Greenville Healthcare System postdoctoral program.

References

Alexe, S., Blackstone, E., Hammer, P. L., Ishwaran, H., Lauer, M. S., & Snader, C. E. P. (2003). Coronary risk prediction by logical analysis of data. *Annals of Operations Research*, *119*(1–4), 15–42.

Almdal, T., Scharling, H., Jensen, J. S., & Vestergaard, H. (2004). The independent effect of type 2 diabetes mellitus on ischemic heart disease, stroke, and death: A population-based study of 13,000 men and women with 20 years of follow-up. *Archives of Internal Medicine*, *164*(13), 1422–1426.

American Diabetes Association. (2006). Diagnosis and classification of diabetes mellitus. *Diabetes Care, 29*(Supplement 1), S43–S48.

American Diabetes Association. (2015). Classification and diagnosis of diabetes. *Diabetes Care, 38*(Supplement 1), S8–S16.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29.

Blanco, R., Larrañaga, P., Inza, I., & Sierra, B. (2004). Gene selection for cancer classification using wrapper approaches. *International Journal of Pattern Recognition and Artificial Intelligence*, *18*(08), 1373–1390.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brolin, R. (1996). Gastrointestinal surgery for severe obesity. *Nutrition*, *12*(6), 403–404.

Buchwald, H. (2005). Bariatric surgery for morbid obesity: Health implications for patients, health professionals, and third-party payers. *Journal of the American College of Surgeons*, *200*(4), 593–604.

Buchwald, H., Avidor, Y., Braunwald, E., Jensen, M. D., Pories, W., Fahrbach, K., et al. (2004). Bariatric surgery: A systematic review and meta-analysis. *JAMA*, *292*(14), 1724–1737.

Buchwald, H., Estok, R., Fahrbach, K., Banel, D., Jensen, M. D., Pories, W. J., et al. (2009). Weight and type 2 diabetes after bariatric surgery: Systematic review and meta-analysis. *The American Journal of Medicine*, *122*(3), 248–256.

Cawley, J., & Meyerhoefer, C. (2012). The medical care costs of obesity: An instrumental variables approach. *Journal of Health Economics*, *31*(1), 219–230.

Centers for Disease Control and Prevention. (2011). National diabetes fact sheet: National estimates and general information on diabetes and prediabetes in the United States, 2011. *Atlanta, GA: US department of health and human services, centers for disease control and prevention, 201*(1).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Daousi, C., Casson, I., Gill, G., MacFarlane, I., Wilding, J., & Pinkney, J. (2006). Prevalence of obesity in type 2 diabetes in secondary care: Association with cardiovascular risk factors. *Postgraduate Medical Journal*, *82*(966), 280–284.

Deeba, F., Mohammed, S. K., Bui, F. M., & Wahid, K. A. (2016). An empirical study on the effect of imbalanced data on bleeding detection in endoscopic video. In *2016 IEEE 38th annual international conference of the engineering in medicine and biology society (EMBC)* (pp. 2598–2601). IEEE.

DeMaria, E. J., Portenier, D., & Wolfe, L. (2007). Obesity surgery mortality risk score: Proposal for a clinically useful score to predict mortality risk in patients undergoing gastric bypass. *Surgery for Obesity and Related Diseases*, *3*(2), 134–140.

Fan, Y. J., & Chaovalitwongse, W. A. (2010). Optimizing feature selection to improve medical diagnosis. *Annals of Operations Research*, *174*(1), 169–183.

Freund, Y., & Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory* (pp. 23–37). Springer.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning. Springer series in statistics* (Vol. 1). Berlin: Springer.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(4), 463–484.

Galar, M., Fernández, A., Barrenechea, E., & Herrera, F. (2013). EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recognition*, *46*(12), 3460–3471.

Grundy, S., Barondess, J., Bellegie, N., Fromm, H., Greenway, F., Halsted, C., et al. (1991). Gastrointestinal surgery for severe obesity. *Annals of Internal Medicine*, *115*(12), 956–961.

Gu, Q., Zhu, L., & Cai, Z. (2009). Evaluation measures of the classification performance of imbalanced data sets. In *International symposium on intelligence computation and applications* (pp. 461–471). Springer.

Hall, M. A. (1999). Correlation-based feature selection for machine learning. Ph.D. thesis, The University of Waikato.

Hall, M. A. (2000). Correlation-based feature selection of discrete and numeric class machine learning. University of Waikato, Department of Computer Science.

Inza, I., Larrañaga, P., Blanco, R., & Cerrolaza, A. J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine*, *31*(2), 91–103.

Inza, I., Larrañaga, P., Etxeberria, R., & Sierra, B. (2000). Feature subset selection by bayesian network-based optimization. *Artificial Intelligence*, *123*(1–2), 157–184.

Jirapech-Umpai, T., & Aitken, S. (2005). Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics*, *6*(1), 148.

Johnson, B. L., Blackhurst, D. W., Latham, B. B., Cull, D. L., Bour, E. S., Oliver, T. L., et al. (2013). Bariatric surgery is associated with a reduction in major macrovascular and microvascular complications in moderately to severely obese patients with type 2 diabetes mellitus. *Journal of the American College of Surgeons*, *216*(4), 545–556.

Johnson, R. J., Johnson, B. L., Blackhurst, D. W., Bour, E. S., Cobb, W. S., Carbonell, A. M., et al. (2012). Bariatric surgery is associated with a reduced risk of mortality in morbidly obese patients with a history of major cardiovascular events. *The American Surgeon*, *78*(6), 685–692.

Kannel, W. B., & McGee, D. L. (1979). Diabetes and cardiovascular disease: The Framingham study. *JAMA*, *241*(19), 2035–2038.

Karegowda, A. G., Manjunath, A., & Jayaram, M. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, *2*(2), 271–277.

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, *11*(1), 51.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18–22.

Li, L., Weinberg, C. R., Darden, T. A., & Pedersen, L. G. (2001). Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, *17*(12), 1131–1142.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113–141.

Ogden, C. L., Carroll, M. D., Fryar, C. D., & Flegal, K. M. (2015). Prevalence of obesity among adults and youth: United States, 2011–2014. *NCHS Data Brief*, *219*(219), 1–8.

Ooi, C., & Tan, P. (2003). Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, *19*(1), 37–44.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, *6*(3), 21–45.

Pories, W. J. (2008). Bariatric surgery: Risks and rewards. *The Journal of Clinical Endocrinology and Metabolism*, *93*(11 Supplement 1), s89–s96.

Quinlan, J. R. (2014). *C4.5: Programs for machine learning*. Amsterdam: Elsevier.

Razzaghi, T., Safro, I. (2015). Scalable multilevel support vector machines. In *ICCS* (pp. 2683–2687).

Razzaghi, T., Roderick, O., Safro, I., & Marko, N. (2016). Multilevel weighted support vector machine for classification on healthcare data with missing values. *PLoS ONE*, *11*(5), e0155,119.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1), 1–39.

Roumani, Y. F., May, J. H., Strum, D. P., & Vargas, L. G. (2013). Classifying highly imbalanced ICU data. *Health Care Management Science*, *16*(2), 119–128.

Roumani, Y. F., Roumani, Y., Nwankpa, J. K., & Tanniru, M. (2018). Classifying readmissions to a cardiac intensive care unit. *Annals of Operations Research*, *263*(1–2), 429–451.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517.

Sarker, A., & Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, *53*, 196–207.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227.

Şeref, O., Razzaghi, T., & Xanthopoulos, P. (2017). Weighted relaxed support vector machines. *Annals of Operations Research*, *249*(1–2), 235–271.

Stamler, J., Vaccaro, O., Neaton, J. D., & Wentworth, D. (1993). Diabetes, other risk factors, and 12-yr cardiovascular mortality for men screened in the multiple risk factor intervention trial. *Diabetes Care*, *16*(2), 434–444.

Taft, L., Evans, R. S., Shyu, C., Egger, M., Chawla, N., Mitchell, J., et al. (2009). Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery. *Journal of Biomedical Informatics*, *42*(2), 356–364.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. Burlington: Morgan Kaufmann.

World Health Organization. (2016). Global report on diabetes. World Health Organization.

Xiong, M., Fang, X., & Zhao, J. (2001). Biomarker identification by feature wrappers. *Genome Research*, *11*(11), 1878–1887.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *ICML*, *97*, 412–420.

Zadrozny, B., Langford, J., & Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. In *Third IEEE international conference on data mining, 2003. ICDM 2003* (pp. 435–442). IEEE.

Zheng, B., Zhang, J., Yoon, S. W., Lam, S. S., Khasawneh, M., & Poranki, S. (2015). Predictive modeling of hospital readmissions using metaheuristics and data mining. *Expert Systems with Applications*, *42*(20), 7110–7120.

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

Affiliations

Talayeh Razzaghi[1] · **Ilya Safro**[2] · **Joseph Ewing**[3] · **Ehsan Sadrfaridpour**[2] · **John D. Scott**[4]

[1] Department of Industrial Engineering, EC III, Room 288, MSC 4230, New Mexico State University, P.O. Box 30001, Las Cruces, NM 88003-8001, USA

[2] School of Computing, Clemson University, Clemson, SC, USA

[3] Quality Management Department, Greenville Health System, Greenville, SC, USA

[4] Department of Surgery, Greenville Hospital System University Medical Center, Greenville, SC, USA