Original Article

Healthc Inform Res. 2020 October;26(4):284-294.
https://doi.org/10.4258/hir.2020.26.4.284
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Mortality Prediction from Hospital-Acquired Infections in Trauma Patients Using an Unbalanced Dataset

Mehrdad Karajizadeh[1], Mahdi Nasiri[1], Mahnaz Yadollahi[2], Amir Hussain Zolfaghari[3], Ali Pakdam[1]

[1]School of Management & Information Sciences, Shiraz University of Medical Sciences, Shiraz, Iran
[2]Trauma Research Center, Shahid Rajaee (Emtiaz) Trauma Hospital, Shiraz University of Medical Sciences, Shiraz, Iran
[3]Department of Computer Science, Laurentian University, Sudbury, Canada

Objectives: Machine learning has been widely used to predict diseases, and it is used to derive impressive knowledge in the healthcare domain. Our objective was to predict in-hospital mortality from hospital-acquired infections in trauma patients on an unbalanced dataset. **Methods:** Our study was a cross-sectional analysis on trauma patients with hospital-acquired infections who were admitted to Shiraz Trauma Hospital from March 20, 2017, to March 21, 2018. The study data was obtained from the surveillance hospital infection database. The data included sex, age, mechanism of injury, body region injured, severity score, type of intervention, infection day after admission, and microorganism causes of infections. We developed our mortality prediction model by random under-sampling, random over-sampling, clustering (k-mean)-C5.0, SMOTE-C5.0, ADASYN-C5.5, SMOTE-SVM, ADASYN-SVM, SMOTE-ANN, and ADASYN-ANN among hospital-acquired infections in trauma patients. All mortality predictions were conducted by IBM SPSS Modeler 18. **Results:** We studied 549 individuals with hospital-acquired infections in a trauma hospital in Shiraz during 2017 and 2018. Prediction accuracy before balancing of the dataset was 86.16%. In contrast, the prediction accuracy for the balanced dataset achieved by random under-sampling, random over-sampling, clustering (k-mean)-C5.0, SMOTE-C5.0, ADASYN-C5.5, and SMOTE-SVM was 70.69%, 94.74%, 93.02%, 93.66%, 90.93%, and 100%, respectively. **Conclusions:** Our findings demonstrate that cleaning an unbalanced dataset increases the accuracy of the classification model. Also, predicting mortality by a clustered under-sampling approach was more precise in comparison to random under-sampling and random over-sampling methods.

Keywords: Machine Learning, Mortality, Injuries, Healthcare Associated Infections, Data Mining, Decision Tree, C5.0

Submitted: January 17, 2020, **Revised:** 1st, March 30, 2020; 2nd, June 11, 2020; 3rd, September 17, 2020, **Accepted:** October 23, 2020

Corresponding Author
Mehrdad Karajizadeh
Health Human Resources Research Center, School of Management & Information Sciences, Shiraz University of Medical Sciences, Alley 29, Qasrodasht Ave, Shiraz, Iran. Tel: +98-713-3234-0774, E-mail: Mehrdad.karaji@gmail.com (https://orcid.org/0000-0002-9297-3488)

I. Introduction

Healthcare data mining has been widely used to help predict diseases and extract impressive knowledge [1], and it is commonly applied to detect early progress of diseases. These techniques can be applied to detect cancer, Alzheimer disease, transient ischemic attacks, lung nodules, coating on the tongue, diabetes, hepatitis, traumatic events, polyps, acute pediatric conditions, and Parkinson's disease [2]. Typically, the prediction variable is unbalanced, which means that one class does not have as many records as the other. The largest class is called the majority, and the smallest class is called the minority [3]. Prediction models using unbalanced data are intricate, as long as balanced training sets are required for standard classifiers learning, such as logistic regression, decision tree, support vector machine (SVM), neural networks, and deep learning. Models often underestimate rare classes in terms of unbalanced data, while the overlapping between two classes will happen.

There are many methods to deal with unbalanced learning, such as data level, algorithm-level, and hybrid methods. In data-level methods, researchers modify the training dataset to make it appropriate for a classifier algorithm. For balance distribution, they might generate new objects for the minority group (over-sampling) and remove instances from majority groups (under-sampling). In algorithm-level methods, they tune existing learners to decrease their bias toward the majority groups, while the cost-sensitive approach is the most commonly used algorithm-level method [4]. Our aim is to predict death by applying various methods of balancing to data on hospital-acquired infection among trauma patients. In medical datasets, records in minority classes are often more vital than those of the control class. Hence, it is critical to handle unbalanced data to improve recognition rates, while it is remarkable that the balancing method depends on the context.

Trauma is a leading cause of death worldwide, while these injured patients usually acquire infections during hospitalization [5]. These infections are the principal cause of mortality and extended hospitalization for trauma patients [6]. Moreover, these types of mortality are among the top five causes of death throughout the world [7]. Trauma patients with hospital-acquired infections have a significantly increased risk of mortality, longer stays in the hospital, and increased cost of equipment or services [8,9], resulting in the nosocomial cause of 80% of in-hospital mortality [10].

Although numerous studies have been done on balancing, there has been little research on the prediction of mortality from hospital-acquired infections in trauma patients using a balanced dataset. On the other hand, context, environment, and predictor variables (such as injury severity score and injury body region) affect the prognostic model. A previous study in Shiraz Trauma Center showed that the accuracy of the traditional scoring system for predicting mortality in trauma patients is under 91% [11]. This research is one of the first works on this topic that handles unbalanced data. We compared various method of data balancing to predict death related to hospital-acquired infections in trauma patients based on a real dataset gathered in a tertiary-care teaching trauma hospital in Shiraz, Iran. This study tries to determine the best method to precisely predict the death rate for hospital-acquired infections in trauma patients. Accurate prediction models can provide useful information for decision making to manage hospital-acquired infections as a priority in terms of patient treatment.

The objectives of this study were the following:

(1) Predicting death from hospital-acquired infections in trauma patients in the absence of a balanced dataset (C5.0 and CHAID);
(2) Predicting death from hospital-acquired infection in the trauma patients using a balanced dataset by sampling methods (reduced data set) (C5.0 and CHAID);
(3) Clustering hospital-acquired infections in trauma patients by k-means algorithms;
(4) Predicting death from hospital-acquired infections in trauma patients in each cluster (C5.0 and CHAID);
(5) Predicting death from hospital-acquired infections in trauma patients with SMOTE-C5.0 and ADASYN-C5.0;
(6) Predicting death from hospital-acquired infections in the trauma patients with SMOTE-SVM, ADASYN-SVM, SMOTE-ANN, and ADASYN-ANN.

Many previous studies have attempted to handle unbalanced data [12-14] by adopting various approaches, such as using the right evaluation metrics, resampling the training set (under-sampling, and over-sampling), using K-fold cross-validation appropriately, ensemble different resampled datasets, resampling different ratios, and clustering the frequent class. However, no best model for these problems has been identified, while this strongly relates to techniques, models, and subjects used [2].

In 2013, Roumani et al. [15] indicated that the C5 and SVM algorithms have the highest recall and specificity, respectively, to predict death in an extremely unbalanced ICU dataset. In 2017, Gu et al. [2] reviewed class unbalanced data and provided techniques to balance data, such as data pre-

processing, classification algorithms, and model evaluation. In 2016, Krawczyk [4] reviewed learning methods for unbalanced data and studied various aspects of unbalanced learning, such as classification, clustering, regression, datastream mining, and big data analytics. Further, they directed handling unbalanced data for all domains. Additionally, in 2011, Paoin [16] observed that the accuracy of the C5.0 and naive Bayes algorithms for predicting death is under 40%.

II. Methods

This study was a cross-sectional analysis on trauma patients with hospital-acquired infections who were admitted to Shiraz Trauma Hospital from March 20, 2017, to March 21, 2018. We aimed to classify unbalanced death records from hospital-acquired infections in trauma patients.

For this purpose, we used the cross-industry standard process for data mining (CRISP-DM) to classify highly unbalanced data. CRISP-DM consists of six steps, namely, identifying the problem, understanding the data, preparing the data, modeling, evaluation, and deployment. It could be a cyclical process [17].

Shiraz Trauma Hospital is affiliated with Shiraz University of Medical Sciences, a national university, which collected hospital-acquired infections data for surveillance and prevention of infections. This reporting aims to reduce hospital-acquired infections.

First, the hospital acquired infection records extracted from the mortality infection management database. Next, all features of hospital-acquired infection analysis were done for descriptive statistics: frequency and mean ± standard deviation (SD). Bivariate analysis was performed, and a *p*-value under 0.05 was considered as a significant level. Further, data preprocessing was done to enhance the data mining process using three stages: data selection, cleaning, and transformation.

We set some rules for our inclusion criteria. We included all trauma patients above 15 years old who had sustained hospital-acquired infections who were injured in road traffic accidents (car, motorcycle, and pedestrian accidents), falls, assaults, and gunshots, or had been struck by an object. We excluded admissions for surgical procedure (elective), complications of previous trauma surgeries, patients who had been burned, foreign body injuries, suicides, and sports injuries, and those who referred to another hospital in Shiraz. Note that patients younger than 15 years old were excluded because they were referred to another hospital in Shiraz.

Finally, records of a total of 549 trauma patients with hospital-acquired infections were selected. The values (sex, age, mechanism of injury, body region injured, severity score, type of intervention, infection day after admission, microorganism causes of infections, and outcome) were chosen from this hospital-acquired infection management database.

This substantial clinical database tends to be incomplete, dirty, inaccurate, and inconsistent. Hence, for the preparation step, we removed duplicate records, found missing values, eliminated outliers, and revised inconsistency in the database. We randomly split data into training (70%), testing (20%), and validation (10%) sets. Moreover, on building the decision tree model (CHAID), we stopped when the minimum records in the parent and child branches became 2% and 1%, respectively. In the CHAID algorithm, a *p*-value of at least 0.05 was considered significant.

All data were transformed to an appropriate format for the IBM SPSS Modeler software (IBM, Armonk, NY, USA). Some new features were also derived using other fields. For example, age was calculated by the expiring date and the birthdate. Next, we divided the participants into three age groups based on a previous study: between 15 and 45, between 46 and 64, and above 65 years [18]. Table 1 presents other categorized variables used.

Furthermore, we applied a decision-tree model for classification considering the study of Alonso et al. [19], which showed that decision-tree models are the conventional techniques in mental health. Hence, the C5.0 and CHAID algorithms were applied for classification. For the CHAID algorithm, we also used a chi-square test to decide the condition for splitting [20]. The following objectives were carried out by using the C5.0 and CHAID algorithms:

(1) To predict the death rate from hospital-acquired infections in trauma patients in the absence of a balanced dataset (using C5.0 and CHAID);
(2) To predict the death rate from hospital-acquired infections in trauma patients using a balanced dataset by using sampling methods (reduced dataset, C5.0, and CHAID);
(3) To cluster hospital-acquired infections in trauma patients by k-means algorithm;
(4) To predict the death rate from hospital-acquired infections in trauma patients regarding each cluster (C5.0 and CHAID);
(5) To predict death from hospital-acquired infections in trauma patients by using SMOTE-C5.0 and ADASYN-C5.0;
(6) To predict death from hospital-acquired infections in trauma patients by using SMOTE-SVM, ADASYN-

 https://doi.org/10.4258/hir.2020.26.4.284

Table 1. Detailed information about dataset used in this study

	Data variable name	Measurement	Data variable categories or values	Role	Definition of the data variable
1	Sex	Nominal	0 = Female 1 = Male	Input	The patient's gender
2	Age category	Ordinal	1 = "15–45" 2 = "46–64" 3 = ">=65"	Input	The patient's age at the time of injury
3	Mechanism of injury	Nominal	1 = Car accident 2 = Motorcycle accident 3 = Pedestrian 4 = Assault 5 = falling 6 = Struck by objects	Input	The mechanism (or multiple injury factor) that caused the injury event
4	Injured body region	Nominal	1 = Head and neck 2 = Face 3 = Thorax 4 = Abdomen 5 = Extremities 6 = Multiple injuries	Input	ISS body region
5	Injury Severity Score (ISS) category	Ordinal	1 = "1–8" 2 = "9–15" 3 = ">=16"	Input	ISS was calculated based on the Baker formula. The ISS severity score that reflects the patient's injuries.
6	Ward	Nominal	1 = ICU 2 = General or surgical ward	Input	Ward where detect nosocomial infection
7	Type of invasive intervention	Nominal	1 = Catheter vein 2 = Urinary catheter 3 = Medical ventilator 4 = Tracheostomy 5 = Trachea intubation 6 = Arterial line 7 = Surgery	Input	Type of invasive intervention performed
8	Infected day	Nominal	1 = Infection is less than 21 day 2 = Infection is higher than 22 day	Input	Substation detect infection date from admission date
9	Hospital-acquired infected	Nominal	1 = upper respiratory infection 2 = Urinary tract infection - other UTI 3 = Surgical site infection - SKIN 4 = Bloodstream infection 5 = Pneumonia 6 = Upper respiratory infection - symptomatic UTI 7 = Central nervous system - meningitis 8 = Surgical site infection - surgery took place	Input	Type of hospital-acquired infections
10	Survival status	Nominal	0 = Non-survivors 1 = Survivors	Target	Survival status when patients discharge

ICU: intensive care unit, UTI: urinary tract infection.

Table 2. Bivariate analysis of mortality predictors

	Survivors (n = 464)	Non-survivors (n = 85)	Total (n = 549)	*p*-value
Sex				0.137
Male	386 (85.6)	65 (14.4)	451 (100)	
Female	78 (79.6)	20 (20.4)	98 (100)	
Age (yr)				<0.05
15–45	318 (89.8)	36 (10.2)	354 (100)	
46-64	84 (81.6)	19 (18.4)	103 (100)	
>65	62 (67.4)	30 (32.6)	92 (100)	
Mechanism of injury				<0.05
Car accident	188 (86.2)	30 (13.8)	218 (100)	
Motorcycle accident	117 (88.6)	15 (11.4)	132 (100)	
Pedestrian	61 (82.4)	13 (17.6)	74 (100)	
Gunshot	8 (66.7)	4 (33.3)	12 (100)	
Falling	65 (74.7)	22 (25.3)	87 (100)	
Assault	13 (100)	0 (0)	13 (100)	
Struck by objects	13 (100)	0 (0)	13 (100)	
Injured body region				0.38
Head and neck	183 (84.7)	33 (15.3)	216 (100)	
Face	17 (81)	4 (19)	21 (100)	
Thorax	54 (84.4)	10 (15.6)	64 (100)	
Abdomen	16 (94.1)	1 (5.9)	17 (100)	
Extremities	107 (88.4)	14 (11.6)	121 (100)	
Multiple Injuries	87 (79.1)	23 (20.9)	110 (100)	
Injury Severity Score (n = 492)				0.18
1–8	157 (89.2)	19 (10.8)	176 (100)	
9–15	170 (82.5)	36 (17.5)	206 (100)	
≥16	94 (85.5)	16 (14.5)	110 (100)	
Ward				<0.05
ICU	312 (80.4)	76 (19.6)	388 (100)	
General or surgical ward	152 (94.4)	9 (5.6)	161 (100)	
Type of invasive intervention				
Catheter vein (yes)	86 (89.6)	10 (10.4)	96 (100)	0.13
Urinary catheter (yes)	113 (90.4)	12 (9.6)	125 (100)	<0.05
Medical ventilator (yes)	102 (75)	34 (25)	136 (100)	<0.05
Tracheostomy (yes)	74 (87.1)	11 (12.9)	85 (100)	0.48
Trachea intubation (yes)	14 (70)	6 (30)	20 (100)	0.06
Arterial line (yes)	2 (100)	0 (0)	2 (100)	0.54
Surgery (yes)	74 (88.1)	10 (11.9)	84 (100)	0.32
Infected day				0.51
Infected in less than 21 days after admission	415 (84.9)	74 (15.1)	489 (100)	
Infected in more than 22 days after admission	49 (81.7)	11 (18.3)	60 (100)	

https://doi.org/10.4258/hir.2020.26.4.284

Table 2. Continued

	Survivors (n = 464)	Non-survivors (n = 85)	Total (n = 549)	*p*-value
Hospital-acquired infected				
Upper respiratory infection (yes)	252 (83.7)	49 (16.3)	301 (100)	0.57
Urinary tract infection - other UTI (yes)	90 (85.7)	15 (14.3)	105 (100)	0.70
Surgical site infection - SKIN (yes)	92 (85.2)	16 (14.8)	108 (100)	0.83
Bloodstream infection (yes)	82 (80.4)	20 (19.6)	102 (100)	0.20
Pneumonia (yes)	34 (85)	6 (15)	40 (100)	0.93
Upper respiratory infection - symptomatic UTI (yes)	14 (87.5)	2 (12.5)	16 (100)	0.73
Central nervous system - meningitis (yes)	17 (70.8)	7 (29.2)	24 (100)	<0.05
Surgical site infection - surgery took place (yes)	1 (50)	1 (50)	2 (100)	0.17

Values are presented as number (%).

ICU: intensive care unit, UTI: urinary tract infection.

SVM, SMOTE-ANN, and ADASYN-ANN.

The following tools were used in this study: IBM SPSS Modeler, MS Excel, SPSS, and Python (for running SMOTE and ADASYN).

We calculated the accuracy, precision, and recall for each classifier algorithm to evaluate each model separately. Previous studies found that these metrics were commonly used to assess the performance of prognostic models [21,22]. In addition, the receiver operating characteristic curve is a standard technique for evaluating classifier performance, and the area under the curve (AUC) is another typical metric for a ROC curve. Hence, we measured the AUC in this study [21].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

III. Results

There were 549 individuals who acquired hospital infections in this trauma hospital during the study period from March 2017 to March 2018. In the studied population, 82.1% were male, and 17.9% were female; 64.5% were aged between 15 to 45 years. The total number of patients with hospital-acquired infections who passed away in the hospital was 85 (15.5%), while the remaining 464 (84.5%) survived. Table 2 shows the demographic characteristic of the studied individuals.

In this study, a death prediction model was applied to unbalanced hospital-acquired infection datasets. Mortality was significantly associated with age, gender, ward, urinary catheter, medical ventilator (yes), and central nervous system - meningitis (yes) (all $p < 0.05$). Table 2 depicts the detailed bivariate analysis of mortality predictors of the studied individuals.

We predicted death rates related to hospital-acquired infections for trauma patients based on unbalanced data by using the C5.0 and CHAID algorithms. The prediction accuracy of C5.0 was higher (86.16% vs. 85.16%). The C5.0 precision count for the death class was 17.64%, and for survival was 90.27%. Table 3 displays more details for accuracy, recall, and precision in predicting the possibility of death from these hospital-acquired infections.

On the other hand, considering a balanced dataset, we predicted mortality rates by random-under sampling using the C5.0 and CHAID algorithms. The accuracy for C5.0 was 70.69%, and that for the CHAID algorithm was 61.24%, as shown in Table 4. After we boosted the dataset for oversampling by C5.0 and CHAID, the accuracy reached 94.74% for C5.0; however, it remained relatively low at 79.47% for CHAID (Table 5).

In terms of clustering, we first used k-mean algorithms by setting 5 as the k value. We set the number of clusters (i.e., k = 5) equal to the number of principal infection diagnoses for the majority class (survivor class). Then mortality was predicted separately for each cluster. After all, the mortality prediction accuracy of this model on the clustered data was higher than the previous methods assessed in this study. Table 6 presents the findings in detail.

Table 3. Performance evaluation of death models

Model	Description	AUC	Accuracy (%)	Class	Precision (%)	Recall (%)
CHAID tree	Classification without the balanced data set	0.781	85.16	Survivors	90.27	86.66
				Non-survivors	17.64	62.50
C5.0 tree	Classification without the balanced data set	0.619	86.16	Survivors	99.13	86.46
				Non-survivors	15.29	76.47

AUC: area under the curve.

Table 4. Performance evaluation of death models (random under-sampling)

Model	Description	AUC	Accuracy (%)	Class	Precision (%)	Recall (%)
CHAID tree	Classification using the balanced data set (random under-sampling)	0.709	61.24	Survivors	28.76	80.76
				Non-survivors	94.11	70.79
C5.0 tree	Classification using the balanced data set (random under-sampling)	0.797	70.69	Survivors	61.79	76.38
				Non-survivors	80.00	66.66

AUC: area under the curve.

Table 5. Performance evaluation of death models (random over-sampling)

Model	Description	AUC	Accuracy (%)	Class	Precision (%)	Recall (%)
CHAID tree	Classification with the balanced data set (boost)	0.883	79.47	Survivors	74.35	82.53
				Non-survivors	69.70	76.98
C5.0 tree	Classification with the balanced data set (boost)	0.974	94.74	Survivors	92.02	97.26
				Non-survivors	97.88	92.58

AUC: area under the curve.

Further, we applied SMOTE-C5.0, ADASYN-C5.0, SMOTE-SVM, ADASYN-SVM, SMOTE-ANN, and ADASYN-ANN, while the AUC for death classification using SMOTE-SVM was 1.00 and 0.99 for the ADASYN-SVM algorithm. Table 7 represents the details of calibration of SVM and the ANN algorithm shown in Supplementary Table S1.

To validate the results, we split the data into training (70%), testing (20%), and validation (10%) sets. Table 8 shows the details for the AUC and the accuracy of each approach. The highest validation accuracy was obtained by the k-means algorithm in the clustering approach, followed by the C5.0 algorithm in classification.

IV. Discussion

This research developed models to predict mortality sustained by hospital-acquired infection data set (dead vs. survived) by various methods like over-sampling, under-sampling, and clustered data set using k-means. Next, death predicted by CHAID, C5.0, SMOTE-C5-0, ADASYN-C5.0, SMOTE-SVM, ADASYN-SVM, SMOTE-ANN, and ADASYN-ANN algorithms while each one run separately. Comparing all, the prediction process by clustering method on imbalanced hospital-acquired infection was better than under-sampling and over-sampling methods.

As a part of this study, the best prediction accuracy for mortality from hospital-acquired infection based on an unbalanced dataset was achieved by using the cluster-based algorithm. Alongside our research, regarding cluster-based under-sampling methods, Yen and Lee [23] found that k-means reduces imbalance distribution, and Rahman and Davis [24] noted its significantly better performance on unbalanced cardiovascular data. Likewise, Onan [25] reported the more reliable predictive performance of clustering-based under-sampling methods.

Additionally, our results showed that random over-sampling led to significantly better prediction performance. These results are similar to the findings of Chawla et al. [21], which showed accuracy improvement after the application of a random over-sampling approach to classify a minority class. Nevertheless, random over-sampling approaches are

 https://doi.org/10.4258/hir.2020.26.4.284

Table 6. Performance evaluation for death models on the clustered dataset

Model	Cluster number	AUC	Accuracy (%)	Class	Precision (%)	Recall (%)
CHAID tree	Cluster 1 with alive data and dead data set	0.862	79.19	Survivors	96.40	74.19
				Non-survivors	57.25	92.59
	Cluster 2 with alive data and dead data set	0.961	89.34	Survivors	100	82.64
				Non-survivors	78.35	100
	Cluster 3 with alive data and dead data set	0.987	94.74	Survivors	94.66	94.66
				Non-survivors	95.87	95.87
	Cluster 4 with alive data and dead data set	0.993	97.60	Survivors	97.06	94.28
				Non-survivors	97.89	98.88
	Cluster 5 with alive data and dead data set	0.982	95.05	Survivors	96.59	93.40
				Non-survivors	93.62	96.70
Overall	-	0.962	91.30	Survivors	96.98	83.35
				Non-survivors	82.56	96.78
C5.0 tree	Cluster 1 with alive data and dead data set	0.899	87.25	Survivors	95.80	83.77
				Non-survivors	76.34	93.46
	Cluster 2 with alive data and dead data set	0.944	92.89	Survivors	96.00	90.57
				Non-survivors	89.69	95.60
	Cluster 3 with alive data and dead data set	0.962	94.77	Survivors	96.00	91.14
				Non-survivors	93.81	96.81
	Cluster 4 with alive data and dead data set	0.981	97.60	Survivors	91.18	100
				Non-survivors	100	96.80
	Cluster 5 with alive data and dead data set	0.999	97.80	Survivors	97.72	97.72
				Non-survivors	97.87	97.87
Overall	-	0.965	93.02	Survivors	93.88	88.29
				Non-survivors	90.39	96.04

AUC: area under the curve.

Table 7. Performance evaluation for death models with SMOTE-C5.0 and ADASYN-C5.0

Model	AUC	Accuracy (%)	Class	Precision (%)	Recall (%)
SMOTE-C5.0	0.97	93.66	Survivors	96.35	90.95
			Non-survivors	91.15	96.43
ADASYN-C5.0	0.95	90.93	Survivors	89.60	92.89
			Non-survivors	92.40	88.91
SMOTE-SVM	1.00	100	Survivors	100	100
			Non-survivors	100	100
ADASYN-SVM	0.99	98.57	Survivors	98.74	98.39
			Non-survivors	98.43	98.71
SMOTE-ANN	0.92	91.48	Survivors	86.54	95.74
			Non-survivors	96.27	98.41
ADASYN-ANN	0.97	97.46	Survivors	96.86	98.09
			Non-survivors	98.08	96.83

SVM: support vector machine, ANN: artificial neural network, AUC: area under the curve.

Table 8. Evaluation metrics in training, testing, and validation sets

Model	Evaluation metrics	Training	Testing	Validation
Classification without the balanced data set (with CHAID)	AUC	0.77	0.81	0.76
	Accuracy (%)	82.34	85.57	92.54
Classification without the balanced data set (with C5.0)	AUC	0.59	0.75	0.60
	Accuracy (%)	84.68	88.66	91.04
Classification with balance data set (boost) with CHAID	AUC	0.89	0.87	0.88
	Accuracy (%)	79.11	76.72	82.42
Classification with balance data set (boost) with C5.0	AUC	0.97	0.97	0.97
	Accuracy (%)	92.65	94.71	91.21
Classification with the balanced data set (random under-sampling) with CHAID	AUC	0.64	0.53	0.74
	Accuracy (%)	59.50	48.28	53.57
Classification with the balanced data set (random under-sampling) with C5.0	AUC	0.78	0.80	0.84
	Accuracy (%)	72.07	76.92	73.08
Cluster 1 with alive data and dead data set and classification with C5.5	AUC	0.91	0.82	0.91
	Accuracy (%)	88.29	81.82	87.76
Cluster 2 with alive data and dead data set and classification with C5.5	AUC	0.95	0.91	0.96
	Accuracy (%)	93.94	90.91	90.62
Cluster 3 with alive data and dead data set and classification with C5.5	AUC	0.96	0.95	0.96
	Accuracy (%)	95.76	96.30	88.46
Cluster 4 with alive data and dead data set and classification with C5.5	AUC	0.98	0.98	1.00
	Accuracy (%)	98.86	94.74	94.44
Cluster 5 with alive data and dead data set and classification with C5.5	AUC	0.99	0.99	1.00
	Accuracy (%)	97.54	98.88	100
Cluster 1 with alive data and dead data set and classification with CHAID	AUC	0.88	0.759	0.872
	Accuracy (%)	81.46	72.73	75.51
Cluster 2 with alive data and dead data set and classification with CHAID	AUC	0.955	0.981	0.954
	Accuracy (%)	89.39	93.94	84.38
Cluster 3 with alive data and dead data set and classification with CHAID	AUC	0.982	1.00	0.99
	Accuracy (%)	94.07	96.30	96.15
Cluster 4 with alive data and dead data set and classification with CHAID	AUC	0.99	1.0	1.0
	Accuracy (%)	96.59	100	100
Cluster 5 with alive data and dead data set and classification with CHAID	AUC	0.99	0.95	0.95
	Accuracy (%)	98.36	87.50	89.66
SMOTE-C5.0	AUC	0.98	0.84	0.89
	Accuracy (%)	93.69	79.69	86.52
ADASYN-C5.0	AUC	0.90	0.77	0.69
	Accuracy (%)	86.37	77.16	75.86
SMOTE-SVM	AUC	1.00	0.989	0.98
	Accuracy (%)	100	92.71	94.38
ADASYN-SVM	AUC	0.99	0.89	0.87
	Accuracy (%)	98.57	81.73	80.46

https://doi.org/10.4258/hir.2020.26.4.284

Table 8. Continued

Model	Evaluation metrics	Training	Testing	Validation
SMOTE-ANN	AUC	0.92	0.87	0.86
	Accuracy (%)	91.48	82.29	79.78
ADASYN-ANN	AUC	0.97	0.76	0.61
	Accuracy (%)	97.46	72.59	62.07

AUC: area under the curve.

sometimes inefficient because it can take a long time to prepare unbalanced data [26].

Notably, we compared these three methods for unbalanced data on a hospital-acquired infection dataset; practicing the same methods as future studies on different healthcare data will be valuable. We were interested in doing this comparison; however, the time and resources of the project were limited. Further, external validation using an alternative dataset could improve the assurance of the model; hence, we consider it a limitation in our study.

Original datasets are unclean and sparse. Therefore, the preparation steps for healthcare data take a long time. A further subject to study could be a systematic review of the handling of unbalanced data in healthcare, which is imperative to provide evidence-based approaches.

The results of this study examined two aspects of unbalanced data elaborately, the prognosis of patients with hospital-acquired infection and the need for pre-processing these types of data.

Interestingly, various balancing approaches were applied to handle the imbalance issue for hospital-acquired infection data in the trauma hospital. What stands out in these types of data is that clustered under-sampling performed better than random over-sampling and under-sampling. Overall, the issue of unbalanced data in healthcare remains from prevention to prognosis and follow-up. Hence, we suggest methods for handling unbalanced data in the healthcare domain.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

The authors would like to acknowledge Tiffany Armstrong from Laurentian University of Canada for proofreading and improving the language. The authors also appreciate the contribution of Trauma Research Center members affiliated with the Shiraz University of Medical Science and nosocomial supervising of Shiraz Trauma Hospital and their colleagues for data collection.

ORCID

Mehrdad Karajizadeh (http://orcid.org/0000-0002-9297-3488)
Mahdi Nasiri (http://orcid.org/0000-0001-6216-0817)
Mahnaz Yadollahi (http://orcid.org/0000-0002-6434-0931)
Amir Hussain Zolfaghari (http://orcid.org/0000-0003-2913-1330)
Ali Pakdam (http://orcid.org/0000-0002-2793-2639)

Supplementary Materials

Supplementary materials can be found via https://doi.org/10.4258/hir.2020.26.4.284.

References

1. Lee DG, Ryu KS, Bashir M, Bae JW, Ryu KH. Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction. J Med Syst 2013;37(2):9896.
2. Guo H, Li Y, Shang J, Gu M, Huang Y, Gong B. Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl 2017;73:220-39.
3. Li Y, Guo H, Liu X, Li Y, Li J. Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. Knowl Based Syst 2016;94:88-104.
4. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell 2016;5(4):221-32.
5. Wallace WC, Cinat M, Gornick WB, Lekawa ME, Wilson SE. Nosocomial infections in the surgical intensive care unit: a difference between trauma and surgical patients. Am Surg 1999;65(10):987-90.

6. Burke JP. Infection control: a problem for patient safety. N Engl J Med 2003;348(7):651-6.
7. Anderson RN. Deaths: leading causes for 1999. Hyattsville (MD): National Center for Health Statistics; 2001.
8. Czaja AS, Rivara FP, Wang J, Koepsell T, Nathens AB, Jurkovich GJ, et al. Late outcomes of trauma patients with infections during index hospitalization. J Trauma 2009;67(4):805-14.
9. Glance LG, Stone PW, Mukamel DB, Dick AW. Increases in mortality, length of stay, and cost associated with hospital-acquired infections in trauma patients. Arch Surg 2011;146(7):794-801.
10. Sheng WH, Wang JT, Lin MS, Chang SC. Risk factors affecting in-hospital mortality in patients with nosocomial infections. J Formos Med Assoc 2007;106(2):110-8.
11. Yadollahi M, Ghaedsharaf Z, Jamali K, Niakan MH, Pazhuheian F, Karajizadeh M. The accuracy of GAP and MGAP scoring systems in predicting mortality in trauma: a diagnostic accuracy study. Adv J Emerg Med 2020;4(3):e73.
12. Spelmen VS, Porkodi R. A review on handling imbalanced data. Proceedings of 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT); 2018 Mar 1-3; Coimbatore, India. p. 1-11.
13. Saarela M, Ryynanen OP, Ayramo S. Predicting hospital associated disability from imbalanced data using supervised learning. Artif Intell Med 2019;95:88-95.
14. Klikowski J, Wozniak M. Multi sampling random subspace ensemble for imbalanced data stream classification. In: Burduk R., Kurzynski M., Wozniak M, editors. Progress in computer recognition systems. Cham, Switzerland: Springer; 2019. p. 360-9.
15. Roumani YF, May JH, Strum DP, Vargas LG. Classifying highly imbalanced ICU data. Health Care Manag Sci 2013;16(2):119-28.
16. Paoin W. Lessons learned from data mining of WHO mortality database. Methods Inf Med 2011;50(4):380-5.
17. Wirth R, Hipp J. CRISP-DM: towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining; 2000 Apr 11-13; Manchester, UK. p. 29-39.
18. Bolandparvaz S, Yadollahi M, Abbasi HR, Anvar M. Injury patterns among various age and gender groups of trauma patients in southern Iran: a cross-sectional study. Medicine (Baltimore) 2017;96(41):e7812.
19. Alonso SG, de la Torre-Diez I, Hamrioui S, Lopez-Coronado M, Barreno DC, Nozaleda LM, et al. Data mining algorithms and techniques in mental health: a systematic review. J Med Syst 2018;42(9):161.
20. Lin CL, Fan CL. Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan. J Asian Archit Build Eng 2019;18(6):539-53.
21. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-57.
22. Arisholm E, Briand LC, Johannessen EB. A systematic and comprehensive investigation of methods to build and evaluate fault prediction models. J Syst Softw 2010;83(1):2-17.
23. Yen SJ, Lee YS. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst Appl 2009;36(3):5718-27.
24. Rahman MM, Davis D. Cluster based under-sampling for unbalanced cardiovascular data. Proceedings of the World Congress on Engineering (WCE); 2013 Jul 3-5; London, UK.
25. Onan A. Consensus clustering-based undersampling approach to imbalanced learning. Sci Program 2019;2019:5901087.
26. Tyagi AK, Reddy VK. Performance analysis of undersampling and over-sampling techniques for solving class imbalance problem. Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM); 2019 Feb 26-28; Jaipur, India.

https://doi.org/10.4258/hir.2020.26.4.284