Supervised Learning Algorithm Comparison in Discharge Status Prediction of Trauma Patients: Empirical Evaluation

Zahra Kohzadi (PhD Candidate)[1,2], Ali Mohammad Nickfarjam (PhD)[1,2]*, Zeinab Kohzadi (PhD Candidate)[3], Leila Shokrizadeh Arani (PhD)[1,2], Mehrdad Mahdian (PhD)[4], Felix Holl (PhD)[5,6]

[1]Health Information Management Research Center, Kashan University of Medical Sciences, Kashan, Iran
[2]Department of Health Information Management and Technology, Allied Medical Sciences Faculty, Kashan University of Medical Sciences, Kashan, Iran
[3]Department of Medical Informatics, School of Allied Medical Sciences Shahid Beheshti University of Medical Sciences, Tehran, Iran
[4]Trauma Research Center, Kashan University of Medical Sciences, Kashan, Iran
[5]DigiHealth Institute, Neu-Ulm University of Applied Sciences, Neu-Ulm, Germany
[6]Institute for Medical Information Processing, Biometry, and Epidemiology, Ludwig Maximilian University of Munich, Munich, Germany

ABSTRACT

Background: By analyzing information from trauma centers, hospitals can identify crucial performance indicators that affect budgets and present growth opportunities, potentially leading to lower mortality rates and improved health status indicators.

Objective: This study aims to determine the best-supervised algorithm for diagnosing the discharge status of trauma patients.

Material and Methods: This retrospective study used the data, collected by the Kashan Trauma Registry from March 2018 to February 2019. Several supervised algorithms, including Naive Bayes, Logistic Regression, Support Vector Machine, Random Forest, and K-Nearest Neighbors, have been evaluated for predicting the discharge status of trauma patients. The performance metrics of accuracy, precision, recall, and F-measure were used. The hold-out technique was applied to train the data.

Results: The Random Forest algorithm had the best performance among the other algorithms. The best accuracy, precision, recall, and F-measure for Gini index were 84/2%, 79/7%, 78/3%, and 76.4%, and for information gain were 84.6%, 79.6%, 76.8%, and 76/20%, respectively.

Conclusion: The results of this research showed that the supervised algorithms, with proper parameter settings, can help diagnose the discharge status of trauma patients. In addition, data balancing can help improve the performance of the algorithms. However, this claim cannot be generalized because it depends on the type of algorithm and the values of the parameters.

Keywords

Artificial Intelligence; Supervised Machine Learning; Trauma; Trauma Centers

Introduction

The term “Trauma” is employed to characterize injuries that result in substantial physical and psychological harm [1]. Injuries can result from various causes, such as car accidents, falls, drowning, burns, self-harm, or violence toward oneself or others [1] that some of them (29%) were associated with road incidents [2]. Road fatalities in low-income and developing nations have consistently exceeded those in developed countries, as highlighted in the 2018 global status report on road safety [3].

The variation in mortality rates across different locations and countries also corresponds to disparities in the demographics of individuals most

*Corresponding author:
Ali Mohammad Nickfarjam
Health Information Management Research Center, Kashan University of Medical Sciences, Kashan, Iran
E-mail: nickfarjam-a@kaums.ac.ir

Received: 14 September 2023
Accepted: 25 December 2023

affected while traveling on roads. Globally, pedestrian and cyclist fatalities contribute to 26% of all deaths, while motorized two- and three-wheeled vehicle-related fatalities account for an additional 28%. Car occupants make up 29% of the fatalities, while the remaining 17% are attributed to unidentified road users. In Africa, pedestrians and cyclists constitute the largest proportion of fatalities, accounting for a significant 44% of all deaths. In SouthEast Asia and the Western Pacific, motorized two- and three-wheeled vehicle riders account for the majority of fatalities, comprising 43% and 36% respectively [4].

Conversely, the discharge outcomes of trauma patients involve various factors, including the length of hospitalization, rates of mortality, rates of readmission, and the distinction between discharge to home and discharge to healthcare facilities [5-8].

Through a research investigation conducted in the United States, an examination was made into the discharge practices at trauma centers, revealing the factors that influence post-hospital disposition. The study revealed that patient characteristics, such as race, insurance status, and injury severity, played a predictive role in determining the type of care received after hospitalization. Notably, individuals with self-pay status and Black patients exhibited a diminished likelihood of being discharged to secondary care facilities [6]. In another study, the relationship between discharge destination and the 30-day readmission rate among elderly trauma patients was explored. The findings indicated that being discharged to extended care and inpatient rehabilitation facilities independently posed risk factors for hospital readmissions within this demographic [8].

In 2009, the World Health Organization (WHO) published a comprehensive guide aimed at improving the quality of trauma treatment. The primary objective of this guide was to reduce the mortality rate resulting from trauma in low- and middle-income countries, while drawing inspiration from successful strategies implemented in other regions [9]. Emphasized within this document was the necessity to establish hospital trauma care systems and implement quality assessment programs to ensure the provision of high-quality care. Among the various instruments utilized for quality assessment, the trauma registry was identified as the most crucial [9]. For several decades, trauma registries have played a pivotal role in the trauma systems of high-income countries, with substantial evidence supporting their numerous benefits. These registries have significantly enhanced the methods of record-keeping and are commonly utilized to demonstrate the advantages associated with trauma systems [10].

Machine Learning (ML) plays a crucial role in the healthcare industry, enabling the discovery of new knowledge and the identification of patterns to inform decision-making. This cutting-edge field aims to extract valuable and essential information from vast datasets. Analytical methods are necessary to identify crucial information for decision-making in healthcare data. The application of ML offers several benefits, including disease detection, management, and prevention, as well as cost reduction in medical care. It also assists in formulating efficient healthcare policies, developing patient recommendation systems, and creating health profiles. The healthcare industry generates significant amounts of data, so maintaining accurate patient diagnosis and treatment requires effective database management [11].

The complexity and volume of healthcare data make it challenging to extract meaningful insights about patients' health status. This data encompasses various aspects such as therapy costs, hospitals, medical claims, patients, physicians, and medical history. ML techniques are crucial for analyzing and drawing conclusions from such complex data to improve patient care and management. ML can aid in the classification of patients' disorders, assist in treatment and management, predict hospital admission duration, and maintain accurate management information systems. Current technologies and ML approaches are helping reduce costs and identify factors contributing to diseases [12]. Machine learning techniques have found extensive applications in the healthcare field, including predicting the development of type 2 diabetes [13], diagnosing breast cancer [14], indicating chronic diseases [15], solving multi-object fusion detection problems in e-healthcare [16], analyzing COVID-19 through clustering algorithms [17], evaluating healthcare facility performance [18], and enhancing mutual privacy in healthcare IoT systems through clustering strategies [19].

Machine learning has also been extensively utilized in the field of trauma and injury. For example, Roberta et al. conducted a study comparing different machine learning algorithms and a classical linear regression model to evaluate traumatic brain injury patients at different time points [20].

Jen Kuo et al. aimed to develop and validate machine learning models for predicting the mortality of hospitalized motorcycle riders using logistic regression, support vector machine, and decision tree analyses [21]. Fen et al. compared the predictive abilities of twenty-two machine learning models with a logistic regression model, using performance measures such as ROC, AUC, accuracy, F-score, precision, recall, and decision curve analysis [22]. Another research focused on evaluating diagnostic accuracy for traumatic brain injury in elderly patients using various machine learning algorithms [23]. Rau et al. predicted patient deaths using logistic regression, support vector machine, decision tree, naive Bayes, and artificial neural network models [24].

Machine learning has diverse applications in trauma registry systems, encompassing patient classification, predictive modeling, data analysis, and system integration. For instance, machine learning models have demonstrated high prognostic performance and medical validity in predicting recovery post-trauma [25]. These models have also been utilized to anticipate blood product transfusion needs in pediatric patients undergoing craniofacial surgery [26]. Furthermore, machine learning has been applied to identify geospatial and structural factors influencing youth violence [27] and predict the risk of prolonged mechanical ventilation for patients with traumatic brain injury [28]. Finally, machine learning has been applied to surgical imaging for diagnosing and treating spine disorders [29].

Research indicates that supervised algorithms are increasingly being used in trauma treatment; despite these studies, no conclusive evidence has been found.

The algorithms that showed acceptable performance were logistic regression, SVM, and random forest. This study focused on supervised methods to diagnose the discharge status of trauma patients because Kashan Trauma Registry data are local, and machine learning has not been applied to them.

The current study will address these questions at the conclusion of this study:

Among the different algorithms used to classify trauma patients, which one performed better at predicting their discharge status? How does data balancing affect algorithm performance? Which of the following indicators of accuracy, precision, recall, and F-measure is most influenced by data balancing? This paper is arranged in the following manner: The second section will introduce the dataset and algorithm and discuss the methodology.

The model evaluation is discussed in Section 3. A detailed description of the experimental procedures and findings can be found in Section 4. Section 5 presents the research results and discusses its main goals. The conclusion is stated at the end of the paper.

Material and Methods

It is a retrospective study of patients who have received trauma treatment at the Kashan

Trauma Center. A specific time frame of patient treatment at the Kashan Trauma Center is scrutinized in this study. In order to improve trauma care, supervised algorithms are employed to analyze the discharging status of trauma patients, distinguishing between those who have improved and those who haven't.

Dataset

Data from the Kashan Trauma Registry was collected between March 2018 and February 2019. We removed noisy data and outliers after data collection. We excluded missing data, split the data, and then balanced these classes based on SMOTE (Figure 1). A total of 3,930 records were obtained after preprocessing the data. We categorized features numerically and categorically (Table 1). The class label was based on the patient's discharge status, which was either improved (2,642) or non-improved (1,288).

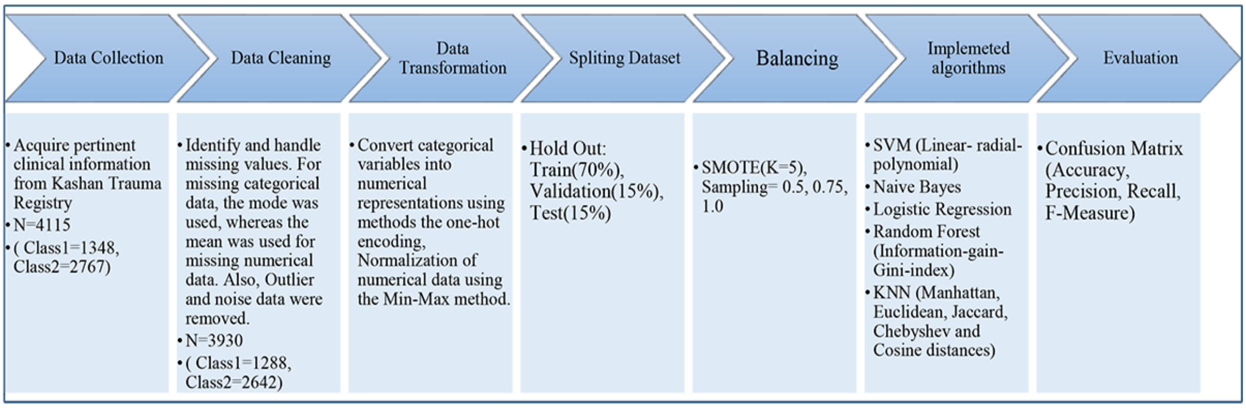
Data Collection
Data Cleaning
Data Transformation
Spliting Dataset
Balancing
Implemeted algorithms
Evaluation
• Acquire pertinent clinical information from Kashan Trauma Registry
• N=4115
• (Class1=1348, Class2=2767)
• Identify and handle missing values. For missing categorical data, the mode was used, whereas the mean was used for missing numerical data. Also, Outlier and noise data were removed.
• N=3930
• (Class1=1288, Class2=2642)
• Convert categorical variables into numerical representations using methods the one-hot encoding, Normalization of numerical data using the Min-Max method.
• Hold Out: Train(70%), Validation(15%), Test(15%)
• SMOTE(K=5), Sampling= 0.5, 0.75, 1.0
• SVM (Linear- radial- polynomial)
• Naive Bayes
• Logistic Regression
• Random Forest (Information-gain- Gini-index)
• KNN (Manhattan, Euclidean, Jaccard, Chebyshev and Cosine distances)
• Confusion Matrix (Accuracy, Precision, Recall, F-Measure)

Figure 1: Implementation Framework (SMOTE: Synthetic Minority Over-sampling Technique, SVM: Support Vector Machine, KNN: K-Nearest Neighbors)

Table 1: Trauma dataset features after preprocessing

Numerical variables	- Age - Total expenditures - The number of days admitted
Categorical variables	- Place birth - Type of insurance - Sex - Occupation - Education - Type of conveyance carrying to emergency - ICD-injuries - ICD-external causes - State of discharge

ICD: International Classification of Diseases

Algorithm selection

During the learning step, a classification model is built, followed by the classification step, a two-part procedure for classifying data (using the model to predict class labels). In the first stage, a classifier is built to describe a preset set of data classes or concepts. During the training phase of the learning stage, a classification algorithm creates the classifier by "learning from" a training set of database tuples and their associated class labels. Since each training tuple has a class label, this process is also known as supervised learning; As a result, the classifier's learning is "supervised" since it knows to which class each training tuple belongs [30, 31]. The following are some descriptions of supervised algorithms used in this article.

SVM

Support Vector Machine (SVM) is a supervised machine learning model that can be utilized for both regression and classification tasks. A key advantage of SVMs is their use of kernels, mathematical functions that project input data into higher-dimensional feature spaces to facilitate separation between classes [32-34]. This projection into hyperspace enables SVM to construct optimal separating hyperplanes between data points of different class labels, improving generalizability and classification accuracy. By effectively separating complex and nonlinear data, SVMs can generate robust predictive models.

KNN

The k-nearest neighbors (KNN) algorithm is a nonparametric, supervised machine learning technique used for both classification and regression predictive modeling. Unlike parametric algorithms, KNN does not make assumptions about the underlying statistical distribution of the data. A key hyperparameter, k, must be predefined to specify the number of neighbors examined during prediction. To determine neighbors, KNN utilizes distance metrics to quantify the similarity between data points, with common choices being Euclidean distance, Manhattan distance, cosine similarity, and Jaccard distance [33, 35]. A key advantage of KNN is that it adapts to the local structure of the data, classifying new points based on their proximity to points in the training set.

Naive Bayes

The Naive Bayes classifier is a probabilistic machine learning algorithm for categorical data predicated on Bayes' theorem. It relies on the simplifying assumption that predictor variables are conditionally independent given the class label - the "naive" conditional independence assumption. Despite this oversimplification, Naive Bayes has demonstrated high predictive performance, especially when dealing with high-dimensional feature spaces [34, 36]. As it requires relatively few training data to estimate parameters, Naive Bayes is an efficient and effective supervised learning technique for multivariate classification tasks. The algorithm calculates posterior probabilities for each class and assigns new data points to the most probable class. Though naive, this Bayesian approach has proven surprisingly robust and applicable across multiple domains.

Logistic Regression

Logistic regression is a statistical technique for modeling the relationship between a categorical dependent variable and one or more independent predictor variables, which may be continuous or categorical [37, 38]. It estimates the probability of particular outcomes, most often binary, based on logistic functions of the linear predictor. Logistic regression facilitates explanatory modeling and prediction of categorical response variables.

Random Forest

Random forest is an ensemble machine learning algorithm that can be utilized for both regression and classification tasks. It operates by constructing a multitude of decision trees during training and aggregating their individual predictions, thereby improving predictive performance and reducing overfitting compared to single decision tree models [34, 39, 40]. To determine the optimal feature split at each node when building individual trees, random forest employs metrics such as information gain, Gini impurity, and gain ratio. By training each tree on a random subset of features and data points, the resulting forest model incorporates diversity while capitalizing on averaging to enhance generalization capability. The algorithm's combination of bagging and random feature selection yields robust and accurate predictions.

SMOTE

SMOTE is a data preprocessing approach that handles class imbalance in machine learning datasets where one class is underrepresented compared to others [41, 42]. Algorithms can struggle to adequately learn patterns and properties of the minority class due to insufficient instances. To mitigate this, SMOTE synthetically generates new minority class examples by interpolating between existing minority data points in feature space [42, 43]. Augmenting the minority class via oversampling improves class balance and enhances model performance on the rare class. By compensating for skewed class distributions, SMOTE facilitates more robust learning from imbalanced data.

Implemented framework

Various supervised machine learning models were implemented for binary patient diagnosis, including support vector machines with linear, radial basis function, and polynomial kernels; k-nearest neighbors' algorithm with Manhattan, Euclidean, Jaccard, Chebyshev, and cosine distance metrics and varying k values; Naive Bayes; logistic regression; and random forests with information gain, gain ratio, and Gini impurity criteria. Algorithms were implemented using Python and Jupyter Notebook.

The dataset was partitioned into 70% training, 15% testing, and 15% validation sets using the hold-out method to enable robust model development, rigorous evaluation, and reliable performance validation. To address the class imbalance in the training data, the SMOTE was applied at sampling rates of 0.5, 0.75, and 1.0 to synthetically generate additional minority class instances.

Evaluation

A confusion matrix is a summarizing the performance of a classification machine learning model by tabulating its True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) predictions [44, 45]. The confusion matrix facilitates the evaluation of key classification metrics including accuracy, precision, recall, and F1-score.

The four confusion matrix categories are [44, 45]:

TP: Correctly predicted non-improved instances

TN: Correctly predicted improved instances

FP: Incorrectly predicted as non-improved when improved

FN: Incorrectly predicted as improved when non-improved

By condensing results into a confusion matrix, model performance on binary classification tasks can be visualized and quantified.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{1}$$

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

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

$$\text{F} - \text{measure} = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{4}$$

A confusion matrix with high accuracy indicates the model correctly classified a substantial proportion of samples [34, 44, 45]. However, accuracy alone can be misleading, especially with imbalanced datasets where one class predominates. While high accuracy is desirable in a confusion matrix, additional metrics should be examined for comprehensive model evaluation. High recall signifies the model correctly identified most actual positives, while high precision indicates few false positives were produced. A high F1 score demonstrates proficiency in both precision and recall, balanced by the F1 measure. Consequently, accuracy, precision, recall, and F1 score were calculated to evaluate the performance of the implemented algorithms. Though accuracy provides an overall measure of correct predictions, precision, and recall offer deeper insight into positive and negative classification capabilities on imbalanced data. The F1 score synthesizes precision and recall into a singular metric, facilitating model selection and performance benchmarking.

Therefore, we calculated the algorithms' accuracy, precision, recall, and F-measure.

Results

Table 2 presents the dataset partition sizes for training, validation, and testing based on the 70/15/15 hold-out split. This allocated 2,751 records for training, 589 for validation, and 590 for testing. After SMOTE oversampling of the training set at rates of 50%, 75%, and 100%, the validation and testing partitions remained unchanged while the training set increased in size as shown in Table 2. Oversampling enabled compensation for class imbalance in the training data only, while preserving untouched validation and test sets for unbiased model evaluation.

The results of executing the algorithms on the test and validation data using SMOTE are shown in Tables 3 and 4. Among the KNN models, K=10 yielded the best performance.

The validation data (Table 3) showed SVM achieved a maximum accuracy of 81.5% with the linear kernel and SMOTE 100%. The linear kernel with SMOTE 75% also yielded the highest precision of 81.5%. Additionally, the linear kernel coupled with SMOTE 100% produced the top recall of 64.8% and the F1 score of 68.4%.

For KNN, Euclidean and Manhattan distances attained a peak accuracy of 80% without SMOTE. Manhattan distance without SMOTE also achieved the highest precision of 79.1%. SMOTE 100% enabled the maximum recall of 78% based on the Jaccard index. An F1 value of 63.6% was obtained using Manhattan distance and SMOTE 75%.

With Naive Bayes, the best accuracy of 65.9% and precision of 46.2% resulted from no SMOTE and SMOTE 50%. SMOTE 100% generated the highest recall of 78% and F1 score of 62.1%.

Logistic regression achieved its maximum accuracy of 81% using SMOTE 50%. The same configuration yielded the top precision of 79.3%. SMOTE 100% produced the highest recall of 76.4% and F1 of 67.4%

Random forest attained its best accuracy of 86.2% and precision of 82.2% with the Gini index and SMOTE 50%. Top recall of 76.4% was seen with SMOTE 100% and 75% for Gini and Information indexes. The maximum F1 of 77% occurred using the Information index and SMOTE 75%.

The test data results (Table 4) showed SVM achieved its highest accuracy of 80.8%, recall of 67.2%, and F1 of 70.2% using the linear kernel and SMOTE 100%. The linear kernel with SMOTE 50% yielded the maximum precision of 85.8%.

In KNN, cosine distance with SMOTE 50% produced the top accuracy of 79.2%. The highest precision was 74.5% without SMOTE for Euclidean distance and with SMOTE 50% for Manhattan distance. SMOTE 100% enabled

Table 2: Partitioning of data before and after balancing

Balancing	Split dataset	The number of class 1 records	The number of class 2 records
Before Balancing	Train	908	1843
	Validation	182	407
	Test	198	392
After Balancing	Train_resampled (100%)	1843	1843
	Train_resampled (75%)	1382	1843
	Train_resampled (50%)	921	1843

Table 3: Performance of algorithms based on the Confusion matrix criteria for different balances (Validation dataset)

Criteria	Balancing (%)	SVM			KNN					NB (%)	LR (%)	RF	
		Linear (%)	Polynomial (%)	RBF (%)	Euclidean (%)	Manhattan (%)	Chebyshev (%)	Cosine (%)	Jaccard (%)			Gini (%)	Information-Gain (%)
Accuracy	SMOTE (100)	81/5	72/5	70/3	73/9	74/5	72/0	72/8	70/6	64/2	78/6	84/4	85/6
	SMOTE (75)	83/0	72/0	71/5	76/9	77/4	72/3	75/9	73/2	64/7	78/8	85/6	85/9
	SMOTE (50)	80/0	73/3	73/5	79/6	79/6	72/2	78/6	74/2	65/9	81/0	86/2	85/6
	SOMTE (0)	79/3	73/3	73/3	80/0	80/0	72/2	78/8	72/5	65/9	80/6	85/6	85/9
Precision	SMOTE (100)	72/4	55/4	51/7	55/9	57/0	59/6	54/7	51/6	45/3	63/7	73/9	77/1
	SMOTE (75)	81/5	55/1	54/4	62/1	63/4	63/0	60/3	55/7	45/4	65/7	78/0	77/7
	SMOTE (50)	81/4	59/8	60/8	76/7	77/7	70/5	72/6	59/3	46/2	79/2	82/2	81/7
	SOMTE (0)	80/6	59/8	60/7	78/1	79/1	70/5	73/2	57/5	46/2	79/3	79/8	81/9
Recall	SMOTE (100)	64/8	56/6	57/7	73/1	71/4	29/1	70/3	78/0	76/4	71/4	76/4	75/8
	SMOTE (75)	58/2	50/5	47/3	64/8	63/7	25/3	64/3	64/8	70/9	65/4	74/2	76/4
	SMOTE (50)	45/6	41/8	40/1	48/9	47/8	17/0	49/5	52/7	63/7	52/2	70/9	68/7
	SOMTE (0)	43/4	41/8	39/0	48/9	47/8	17/0	49/5	42/3	63/7	50/5	71/4	69/8
F-Measure	SMOTE (100)	68/4	56/0	54/5	63/3	63/4	39/1	61/5	62/1	56/9	67/4	75/1	76/5
	SMOTE (75)	67/9	52/7	50/6	63/4	63/6	36/1	62/2	59/9	55/4	65/6	76/1	77/0
	SMOTE (50)	58/5	49/2	48/3	59/7	59/2	27/4	58/8	55/8	53/6	62/9	76/1	74/6
	SOMTE (0)	56/4	49/2	47/5	60/1	59/6	27/4	59/0	48/7	53/6	61/7	75/4	75/4

SVM: Support Vector Machine, KNN: K-Nearest Neighbors, NB: Naive Bayes, LR: Logistic Regression, RF: Random Forest, SMOTE: Synthetic Minority Over-sampling Technique, RBF: Radial Basis Function

Table 4: Performance of algorithms based on the Confusion matrix criteria for different balances (Test dataset)

Criteria	Balancing (%)	SVM			KNN					NB (%)	LR (%)	RF	
		Linear (%)	Polynomial (%)	RBF (%)	Euclidean (%)	Manhattan (%)	Chebyshev (%)	Cosine (%)	Jaccard (%)			Gini (%)	Information-Gain (%)
Accuracy	SMOTE (100)	80/8	72/2	71/9	73/2	73/6	71/7	72/5	70/2	62/0	77/5	83/7	82/4
	SMOTE (75)	80/0	72/2	71/5	75/9	76/3	72/5	75/9	70/7	62/9	78/8	83/9	83/4
	SMOTE (50)	80/2	71/5	72/2	78/3	78/5	71/5	79/2	73/6	64/4	77/5	84/2	84/6
	SOMTE (0)	79/8	72/0	72/0	78/5	78/6	71/5	78/8	72/4	64/7	77/3	84/2	84/6
Precision	SMOTE (100)	73/5	57/5	57/3	58/5	58/8	66/7	57/6	53/9	46/0	64/8	74/5	72/4
	SMOTE (75)	77/0	58/7	58/1	64/0	64/5	73/1	63/9	55/4	46/6	68/9	77/0	75/5
	SMOTE (50)	85/8	60/0	62/3	74/0	74/5	74/2	74/2	60/5	47/9	73/0	79/7	78/9
	SOMTE (0)	85/6	61/1	62/0	74/5	75/0	74/2	73/9	60/9	48/2	72/9	79/7	79/6
Recall	SMOTE (100)	67/2	66/2	63/6	69/7	70/7	31/3	68/7	77/3	75/8	71/7	78/3	76/8
	SMOTE (75)	57/6	58/1	54/5	64/6	65/2	28/8	65/2	65/2	72/7	67/2	74/2	74/7
	SMOTE (50)	49/0	45/5	43/4	54/5	54/5	23/2	58/1	61/1	67/7	52/0	71/2	73/7
	SOMTE (0)	48/0	46/0	42/9	54/5	54/5	23/2	57/1	49/5	67/7	51/5	71/2	72/7
F-Measure	SMOTE (100)	70/2	61/5	60/3	63/6	64/2	42/6	62/7	63/5	57/3	68/1	76/4	74/5
	SMOTE (75)	65/9	58/4	56/3	64/3	64/8	41/3	64/5	59/9	56/8	68/0	75/6	75/1
	SMOTE (50)	62/4	51/7	51/2	62/8	63/0	35/4	65/2	60/8	56/1	60/8	75/2	76/2
	SOMTE (0)	61/5	52/4	50/7	63/0	63/2	35/4	64/4	54/6	56/3	60/4	75/2	76/0

SVM: Support Vector Machine, KNN: K-Nearest Neighbors, NB: Naive Bayes, LR: Logistic Regression, RF: Random Forest, SMOTE: Synthetic Minority Over-sampling Technique, RBF: Radial Basis Function the maximum recall of 77.3% per the Jaccard index. Cosine distance with SMOTE 50% achieved the highest F1 of 65.2%.

With Naive Bayes, no SMOTE yielded the best accuracy of 64.7% and precision of 48.2%. SMOTE 100% generated a peak recall of 75.8% and F1 of 57.3%.

For logistic regression, SMOTE 75% achieved the maximum accuracy of 78.8%. SMOTE 50% yielded the highest precision of 73%. Top recall of 71.7% and F1 of 68.1% resulted from SMOTE 100%.

In random forest, an equal high accuracy of 84.6% occurred with and without SMOTE 50% using the information gain index. The Gini index attained its best precision of 79.7% with SMOTE 50%. Additionally, SMOTE 100% enabled the maximum recall of 78.3% and F1 of 76.4% for the Gini index.

Discussion

Trauma and injury registration encompasses the collection of prehospital data and demographic information pertaining to the occurrence of the injury. The World Health Organization has recommended the utilization of this data for the purpose of effectively managing these patients and enhancing the standard of care provided to them [9]. The quantity of data in our progressively digitalized world is experiencing exponential growth, and big data analytics represents both a burgeoning trend and a prominent area of study. The algorithms employed in machine learning grant access to analyses, enabling the detection and prediction of disease existence, as well as aiding medical professionals in decision-making by facilitating early disease identification and appropriate therapy selection.

Based on the outcome of the present study, the optimal outcomes were observed with SVM, Random Forest (depth=10), and KNN algorithms in which linear kernels were used, along with the Gini index and Information Gain, as well as the Euclidean and Manhattan distances with k set to 10.

In the majority of algorithms, SMOTE with a 50% oversampling rate yielded higher accuracy compared to SMOTE with a 75% oversampling rate and SMOTAE with a 100% oversampling rate. The precision metric showed suboptimal performance with SMOTAE (75%) and SMOTE (100%) in most algorithms. Furthermore, recall, and F-score exhibited an upward trend across most algorithms as the number of balanced records increased.

Nevertheless, it cannot be definitively concluded that SMOTE had a uniformly positive or negative impact on all indicators simultaneously. In certain algorithms, the application of SMOTE appeared to be necessary, while in others, better results were achieved without utilizing SMOTE.

A notable finding in our study was that Naïve Bayes was the algorithm with the weakest performance, whereas Random Forest was the algorithm with the best performance.

In the study conducted by Ruschetta et al. [20] the performance of SVM, KNN, NB, DT algorithms, and an ensemble machine-learning approach was compared individually. The results indicated that the NB algorithm exhibited the poorest performance when a two-class outcome (positive or negative) was employed. Similarly, in the current study, NB was also among the algorithms that demonstrated relatively inferior performance.

The machine learning techniques can be utilized to predict the mortality of motorcycle riders with a reasonable level of accuracy [21]. By integrating a machine learning model, particularly the SVM algorithm, into the trauma system, it may be possible to identify high-risk patients and guide clinical staff towards the most suitable interventions. In the current study, the SVM algorithm with the linear kernel exhibited satisfactory performance.

In the study conducted by Fen et al. it was found that the twenty-two machine learning models selected for outcome prediction in patients with Severe Traumatic Brain Injury

(STBI) exhibited capabilities comparable to the traditional Logistic Regression (LR) model. Notably, the cubic SVM, quadratic SVM, and linear SVM models outperformed LR in terms of performance [22]. In the present study, SVM with a linear kernel was identified as the SVM algorithm with the highest performance. However, the random forest algorithm (using the Gini-Index) demonstrated the best overall performance among all the algorithms tested, although the results obtained with logistic regression were also deemed acceptable.

According to the findings of Abujaber et al. [46], the performance of the SVM algorithm surpassed that of traditional classical models employing conventional multivariate analytical approaches when predicting mortality in patients with TBI. In the current study, although the SVM (linear) algorithm exhibited relatively good performance, it was not the top-performing algorithm.

Similar to the present study, the random forest algorithm showed the best performance, and Logistic Regression yielded acceptable results. In a study by Wang et al. [23], it was reported that prognostication tools utilizing Adaboost, Random Forest, and Logistic Regression algorithms proved beneficial for physicians in assessing the risk of poor outcomes in geriatric patients with TBI and in guiding the selection of personalized therapeutic options.

According to the findings of Matsuo et al. [47], both the Random Forest and Ridge Regression algorithms demonstrated the highest performance in predicting poor in-hospital outcomes and mortality in cases of TBI. Their research indicates that modern machine learning techniques can effectively predict the occurrence of TBI. Similarly, in the current study, the random forest algorithm was identified as the best-performing algorithm among the ones tested.

According to the conclusions drawn from this study, equalizing class features can effectively improve the performance of machine learning algorithms. However, it is important to note that the choice of algorithm, its parameters, and the quantity of added samples can directly impact the algorithm's performance. Therefore, relying solely on accuracy values in scenarios with imbalanced data may not be feasible. The findings of this research suggest the potential use of supervised algorithms for predicting the discharge status of trauma patients.

Despite the advantages of this study, there are some limitations to consider. First, the data used in the study is retrospective, and it was not possible to access the paper records to verify the quality of the electronic data. Additionally, in future research, alternative classification methods with different parameters, ensemble learning techniques, and clustering approaches could be explored to improve the diagnosis of discharge status for trauma patients.

Conclusion

The registration of health data in health systems can benefit from the application of machine learning techniques, which can help health stakeholders uncover hidden knowledge in the data and support them in decision-making and health prediction. While supervised algorithms are valuable in diagnosing the discharge status of trauma patients, the impact of data balancing on accuracy measures such as Precision, Recall, and F-measure varies across different algorithms. These measures do not consistently show a trend of increase or decrease. Therefore, optimizing the performance of algorithms requires appropriate parameter settings. Balancing imbalanced data may improve algorithmic performance, but it is important to note that the effectiveness of this approach depends on the specific algorithm and the parameter values assigned to it. In summary, the success of data balancing in enhancing algorithmic performance hinges on carefully considering algorithm characteristics and configuring parameters accordingly.

Acknowledgment

We acknowledge and appreciate the collaboration of the Trauma Research Center at Kashan University of Medical Sciences. Their contribution has been valuable in conducting this research and advancing our understanding of trauma-related issues.

Authors' Contribution

AM. Nickfarjam, ZA. Kohzadi, M. Mahdian conceived and designed the analysis, collected the data, and contributed to the revision. AM. Nickfarjam, ZA. Kohzadi, ZE. Kohzadi conceived and designed the analysis, contributed data or analysis tools, performed the analysis, wrote, revised, and edited the manuscript, and was involved in the investigation and methodology. AM. Nickfarjam, ZA. Kohzadi, ZE. Kohzadi, F. Holl, L. Shokrizadeh Arani, and M. Mahdian collectively conducted a review of related works in the field of trauma and contributed to writing, editing, and revision tasks. All authors participated in the review and approval of the article.

Ethical Approval

This article is based on a research project that received support from the Kashan University of Medical Sciences, with research approval code 401098 and ethics code IR.KAUMS.NUHEPM.REC.1401.056. The methods employed in this study were carried out in accordance with the applicable guidelines and regulations of the ethical committee at Kashan University of Medical Sciences.

Conflict of Interest

None

References

1. WHO. Injuries and violence. World Health Organization; 2021. Available from: https://www.who.int/news-room/fact-sheets/detail/injuries-and-violence.
2. WHO. Preventing injuries and violence: an overview. World Health Organization; 2022. Available from: https://www.who.int/publications/i/item/9789240047136.
3. WHO. Based on the WHO Global Status Report on Road Safety 2018. World Health Organization; 2018. Available from: https://extranet.who.int/roadsafety/death-on-the-roads/#trends/deaths.
4. WHO. Global Status Report on Alcohol and Health 2018. World Health Organization; 2018. Available from: https://www.who.int/publications/i/item/9789241565639.
5. Narula N, Tsikis S, Jinadasa SP, Parsons CS, Cook CH, Butt B, Odom SR. The Effect of Anticoagulation and Antiplatelet Use in Trauma Patients on Mortality and Length of Stay. *Am Surg*. 2022;**88**(6):1137-45. doi: 10.1177/0003134821989043. PubMed PMID: 33522831.
6. Elkbuli A, Sutherland M, Gargano T, Kinslow K, Liu H, McKenney M, Ang D. Race and Insurance Status Disparities in Post-discharge Disposition After Hospitalization for Major Trauma. *Am Surg*. 2023;**89**(3):379-89. doi: 10.1177/00031348211029864. PubMed PMID: 34176320.
7. Knauf T, Buecking B, Geiger L, Hack J, Schwenzfeur R, Knobe M, et al. The Predictive Value of the "Identification of Seniors at Risk" Score on Mortality, Length of Stay, Mobility and the Destination of Discharge of Geriatric Hip Fracture Patients. *Clin Interv Aging*. 2022;**17**:309-16. doi: 10.2147/CIA.S344689. PubMed PMID: 35386750. PubMed PMCID: PMC8979564.
8. Strosberg DS, Housley BC, Vazquez D, Rushing A, Steinberg S, Jones C. Discharge destination and readmission rates in older trauma patients. *J Surg Res.* 2017;**207**:27-32. doi: 10.1016/j.jss.2016.07.015. PubMed PMID: 27979485.
9. WHO. Guidelines for trauma quality improvement programmes. World Health Organization; 2009. Available from: https://www.who.int/publications/i/item/guidelines-for-trauma-quality-improvement-programmes.
10. Mock C, Nguyen S, Quansah R, Arreola-Risa C, Viradia R, Joshipura M. Evaluation of Trauma Care capabilities in four countries using the WHO-IATSIC Guidelines for Essential Trauma Care. *World J Surg*. 2006;**30**(6):946-56. doi: 10.1007/s00268-005-0768-4. PubMed PMID: 16736320.
11. Varghese DP, Tintu PB. A survey on health data using data mining techniques. *International Research Journal of Engineering and Technology (IRJET)*. 2015;**2**(7):713-20.
12. Ogundele IO, Popoola OL, Oyesola OO, Orija KT. A review on data mining in healthcare. *International Journal of Advanced Research in Computer Engi-*

neering and Technology (IJARCET). 2018;**7**:698-704.

13. Garcia-Carretero R, Vigil-Medina L, Mora-Jimenez I, Soguero-Ruiz C, Barquero-Perez O, Ramos-Lopez J. Use of a K-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population. *Med Biol Eng Comput*. 2020;**58**(5):991-1002. doi: 10.1007/s11517-020-02132-w. PubMed PMID: 32100174.
14. Eedi H, Kolla M. Machine learning approaches for healthcare data analysis. *J Crit Rev*. 2020;**7**(4):806-11. doi: 10.31838/jcr.07.04.149.
15. Induja SN, Raji CG. Computational methods for predicting chronic disease in healthcare communities. In: 2019 International Conference on Data Science and Communication (IconDSC); Bangalore, India: IEEE; 2019. p. 1-6.
16. Ahmad I, Ullah I, Khan WU, Ur Rehman A, Adrees MS, Saleem MQ, et al. Efficient algorithms for E-healthcare to solve multiobject fuse detection problem. *Journal of Healthcare Engineering.* 2021;**2021**:1-6. doi: 10.1155/2021/9500304.
17. Zubair M, Asif Iqbal MD, Shil A, Haque E, Moshiul Hoque M, Sarker IH. An efficient k-means clustering algorithm for analysing covid-19. In: Hybrid Intelligent Systems: International Conference on Hybrid Intelligent Systems (HIS 2020); Springer, Cham; 2020. p. 422-32.
18. Gesicho MB, Were MC, Babic A. Evaluating performance of health care facilities at meeting HIV-indicator reporting requirements in Kenya: an application of K-means clustering algorithm. *BMC Med Inform Decis Mak.* 2021;**21**(1):6. doi: 10.1186/s12911-020-01367-9. PubMed PMID: 33407380. PubMed PMCID: PMC7789797.
19. Guo X, Lin H, Wu Y, Peng M. A new data clustering strategy for enhancing mutual privacy in healthcare IoT systems. *Future Generation Computer Systems*. 2020;**113**:407-17. 10.1016/j.future.2020.07.023.
20. Bruschetta R, Tartarisco G, Lucca LF, Leto E, Ursino M, Tonin P, Pioggia G, Cerasa A. Predicting Outcome of Traumatic Brain Injury: Is Machine Learning the Best Way? *Biomedicines.* 2022;**10**(3):686. doi: 10.3390/biomedicines10030686. PubMed PMID: 35327488. PubMed PMCID: PMC8945356.
21. Kuo PJ, Wu SC, Chien PC, Rau CS, Chen YC, Hsieh HY, Hsieh CH. Derivation and validation of different machine-learning models in mortality prediction of trauma in motorcycle riders: a cross-sectional retrospective study in southern Taiwan. *BMJ Open.* 2018;**8**(1):e018252. doi: 10.1136/bmjopen-2017-018252. PubMed PMID: 29306885. PubMed PMCID: PMC5781097.
22. Feng JZ, Wang Y, Peng J, Sun MW, Zeng J, Jiang H. Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries. *J Crit Care.* 2019;**54**:110-6. doi: 10.1016/j.jcrc.2019.08.010. PubMed PMID: 31408805.
23. Wang R, Zeng X, Long Y, Zhang J, Bo H, He M, Xu J. Prediction of Mortality in Geriatric Traumatic Brain Injury Patients Using Machine Learning Algorithms. *Brain Sci.* 2023;**13**(1):94. doi: 10.3390/brainsci13010094. PubMed PMID: 36672075. PubMed PMCID: PMC9857144.
24. Rau CS, Kuo PJ, Chien PC, Huang CY, Hsieh HY, Hsieh CH. Mortality prediction in patients with isolated moderate and severe traumatic brain injury using machine learning models. *PLoS One.* 2018;**13**(11):e0207192. doi: 10.1371/journal.pone.0207192. PubMed PMID: 30412613. PubMed PMCID: PMC6226171.
25. Stoitsas K, Bahulikar S, De Munter L, De Jongh MAC, Jansen MAC, Jung MM, et al. Clustering of trauma patients based on longitudinal data and the application of machine learning to predict recovery. *Sci Rep.* 2022;**12**(1):16990. doi: 10.1038/s41598-022-21390-2. PubMed PMID: 36216874. PubMed PMCID: PMC9550811.
26. Jalali A, Lonsdale H, Zamora LV, Ahumada L, Nguyen ATH, Rehman M, et al. Machine Learning Applied to Registry Data: Development of a Patient-Specific Prediction Model for Blood Transfusion Requirements During Craniofacial Surgery Using the Pediatric Craniofacial Perioperative Registry Dataset. *Anesth Analg.* 2021;**132**(1):160-71. doi: 10.1213/ANE.0000000000004988. PubMed PMID: 32618624.
27. Doucet JJ, Godat LN, Berndtson AE, Liepert AE, Weaver JL, Smith AM, et al. Youth violence prevention can be enhanced by geospatial analysis of trauma registry data. *J Trauma Acute Care Surg.* 2022;**93**(4):482-7. doi: 10.1097/TA.0000000000003609. PubMed PMID: 35343924.
28. Abujaber A, Fadlalla A, Gammoh D, Abdelrahman H, Mollazehi M, El-Menyar A. Using trauma registry data to predict prolonged mechanical ventilation in patients with traumatic brain injury: Machine learning approach. *PLoS One.* 2020;**15**(7):e0235231. doi: 10.1371/journal.pone.0235231. PubMed PMID: 32639971. PubMed PMCID: PMC7343348.
29. Karandikar P, Massaad E, Hadzipasic M, Kiapour A, Joshi RS, Shankar GM, Shin JH. Machine Learning

Applications of Surgical Imaging for the Diagnosis and Treatment of Spine Disorders: Current State of the Art. *Neurosurgery.* 2022;**90**(4):372-82. doi: 10.1227/NEU.0000000000001853. PubMed PMID: 35107085.

30. Alloghani M, Al-Jumeily D, Mustafina J, Hussain A, Aljaaf AJ. A systematic review on supervised and unsupervised machine learning algorithms for data science. In: Supervised and Unsupervised Learning for Data Science. Springer, Cham; 2020. p: 3-21.

31. Muhammad I, Yan Z. Supervised Machine Learning Approaches: A Survey. *Ictact Journal on Soft Computing*. 2015;**5**(3):946-52. doi: 10.21917/ijsc.2015.0133.

32. Cortes C, Vapnik V. Support vector machine. *Machine learning.* 1995;**20**(3):273-97.

33. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2009.

34. Hackeling G. Mastering Machine Learning with scikit-learn. Packt Publishing Ltd; 2017.

35. Alpaydin E. Introduction to machine learning. 3rd ed: The MIT Press; 2020.

36. Brownlee J. Probability for machine learning: Discover how to harness uncertainty with Python. Machine Learning Mastery; 2019.

37. Pampel FC. Logistic regression: A primer. Second ed. Sage; 2020.

38. Menard S. Logistic regression: From introductory to advanced concepts and applications. Sage; 2010.

39. Breiman L. Random forests. *Machine learning*. 2001;**45**:5-32. doi: 10.1023/A:1010933404324.

40. Cutler A, Cutler DR, Stevens JR. Random forests. In: Ensemble machine learning: Methods applications. New York, NY: Springer; 2012. p. 157-75.

41. Kaur G, Kaur V, Sharma Y, Bansal V. Analyzing various Machine Learning Algorithms with SMOTE and ADASYN for Image Classification having Imbalanced Data. In: International Conference on Current Development in Engineering and Technology (CCET); Bhopal, India: IEEE; 2022. p. 1-7.

42. Fernández A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research.* 2018;**61**:863-905. doi: 10.1613/jair.1.11192.

43. Dowlagar S, Mamidi R. DepressionOne@ LT-EDI-ACL2022: Using Machine Learning with SMOTE and Random UnderSampling to Detect Signs of Depression on Social Media Text. In: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion; LTEDI; 2022. p. 301-5.

44. Bishop CM, Nasrabadi NM. Pattern recognition and machine learning. New York: Springer; 2006.

45. Witten IH, Frank E. Data mining: practical machine learning tools and techniques. Elsevier; 2002.

46. Abujaber A, Fadlalla A, Gammoh D, Abdelrahman H, Mollazehi M, El-Menyar A. Prediction of in-hospital mortality in patients with post traumatic brain injury using National Trauma Registry and Machine Learning Approach. *Scand J Trauma Resusc Emerg Med*. 2020;**28**(1):44. doi: 10.1186/s13049-020-00738-5. PubMed PMID: 32460867. PubMed PMCID: PMC7251921.

47. Matsuo K, Aihara H, Nakai T, Morishita A, Tohma Y, Kohmura E. Machine Learning to Predict In-Hospital Morbidity and Mortality after Traumatic Brain Injury. *J Neurotrauma.* 2020;**37**(1):202-10. doi: 10.1089/neu.2018.6276. PubMed PMID: 31359814.