Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2022;14(3):168-79

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2022-88932

TwoClsBalancer: An Interactive Web Application for Handling the Class Imbalance Problem Based on Machine Learning

TwoClsBalancer: Sınıf Dengesizliği Problemi İçin Makine Öğrenmesine Dayalı Etkileşimli Bir Web Uygulaması

Ahmet Kadir ARSLAN[a], Cemil ÇOLAK[a], Mehmet Cengiz ÇOLAK[b]

[a]Department of Biostatistics and Medical Informatics, İnönü University Faculty of Medicine, Malatya, Türkiye
[b]Department of Cardiovascular Surgery, İnönü University Faculty of Medicine, Malatya, Türkiye

ABSTRACT Objective: The main purpose of this research is to develop a novel user-friendly web tool based on machine learning approaches, which applies a variety of techniques to address the class imbalance problem. **Material and Methods:** Shiny, an open-source R package, was used to develop the proposed web tool. The interactive tool can handle the class imbalance problem for binary classification dataset(s) by implementing sampling-based methods. As a clinical application, the dataset retrospectively obtained from the database of the Cardiovascular Surgery Department of Turgut Özal Medical Center, İnönü University, Malatya, Türkiye was used in this web-based software. To overcome the class imbalance problem, sampling-based methods were implemented on the original dataset. After this process, the classification of hypertension in patients with coronary artery disease was achieved by three classification models. **Results:** According to the outputs of the developed web application, the best classification performance was obtained by the support vector machines with radial basis function kernel (SVM-RBF) model after applying the density-based synthetic minority over-sampling technique oversampling method. The accuracy, sensitivity, specificity, precision, f-measure, and g-mean metrics of the relevant model were calculated as 0.99, 0.99, 0.99, 0.95, 0.97, and 0.97, respectively. **Conclusion:** The oversampling methods used in this research indicated a more positive contribution to the classification performance of the models as compared to the undersampling methods. When the undersampling methods were applied, the three classification models did not demonstrate successful classification performance, whereas the SVM-RBF model outperformed the other two models when the oversampling methods were implemented. The designed interactive web application is freely accessible through http://biostatapps.inonu.edu.tr/twoclsbalancer.

Keywords: Classification; coronary artery disease;
hypertension; class imbalance problem;
web-based application

ÖZET Amaç: Bu araştırmanın temel amacı, sınıf dengesizliği sorununu çözmek için çeşitli teknikler uygulayan makine öğrenimi yaklaşımlarına dayalı yeni, kullanıcı dostu bir web aracı geliştirmektir. **Gereç ve Yöntemler:** Açık kaynaklı bir R paketi olan Shiny, önerilen web aracını geliştirmek için kullanıldı. Etkileşimli araç, örneklemeye dayalı yöntemler uygulayarak ikili sınıflandırma veri kümeleri için sınıf dengesizliği sorununu çözebilir. Web tabanlı bu yazılımda, klinik uygulama olarak Malatya İnönü Üniversitesi Turgut Özal Tıp Merkezi Kalp Damar Cerrahisi Anabilim Dalı veri tabanından geriye dönük olarak elde edilen veri seti kullanılmıştır. Sınıf dengesizliği sorununun üstesinden gelmek için orijinal veri seti üzerinde örneklemeye dayalı yöntemler uygulanmıştır. Bu işlemden sonra koroner arter hastalığı olan hastalarda hipertansiyonun sınıflandırılması üç sınıflandırma modeli ile sağlanmıştır. **Bulgular:** Geliştirilen web uygulamasının çıktılarına göre en iyi sınıflandırma performansı, "density-based synthetic minority oversampling technique" aşırı örnekleme yöntemi uygulandıktan sonra radyal tabanlı destek vektör makineleri [support vector machines with radial basis function (SVM-RBF)] modeli ile elde edilmiştir. İlgili modelin doğruluk, duyarlılık, özgüllük, kesinlik, f-ölçümü ve g-ortalama metrikleri sırasıyla 0,99, 0,99, 0,99, 0,95, 0,97 ve 0,97 olarak hesaplanmıştır. **Sonuç:** Bu araştırmada kullanılan aşırı örnekleme yöntemleri, alt örnekleme yöntemlerine kıyasla modellerin sınıflandırma performansına daha olumlu katkı sağlamıştır. Alt örnekleme yöntemleri uygulandığında, 3 sınıflandırma modeli başarılı sınıflandırma performansı göstermezken, aşırı örnekleme yöntemleri uygulandığında SVM-RBF modeli diğer 2 modelden daha iyi performans göstermiştir. Tasarlanan interaktif web uygulamasına http://biostatapps.inonu.edu.tr/twoclsbalancer adresinden ücretsiz olarak erişilebilir.

Anahtar kelimeler: Sınıflandırma; koroner arter hastalığı;
hipertansiyon; sınıf dengesizliği sorunu;
web tabanlı uygulama

Correspondence: Cemil ÇOLAK
Department of Biostatistics and Medical Informatics, İnönü University Faculty of Medicine, Malatya, Türkiye
E-mail: cemilcolak@yahoo.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 14 Feb 2022 ***Received in revised form:*** 05 Sep 2022 ***Accepted:*** 05 Sep 2022 ***Available online:*** 16 Sep 2022

The amount of data produced in the field of medicine is growing daily and with it, the need to store, manage and make available the enormous amount of data generated. These data, which are difficult to store, manage, analyze and have various and complicated forms, are generally called big data.[1] It is becoming increasingly important to use machine learning (ML) techniques to extract patterns from big datasets produced in areas such as healthcare, banking, and telecommunications, and to facilitate prediction and decision support components.[2] Toward this end, one of the most frequently applied ML tasks is classification. Classification is a prediction process that involves assigning the observations that comprise the dataset into previously determined classes within an established framework. It has been common practice, particularly in the healthcare area, to describe diseases using ML approaches and disease-associated risk factors; this system then assists in determining the relative importance for each risk factor. The uneven distribution of the target variables into classes is a primary cause of these newly developing, unwanted consequences. The class imbalance problem, standard classification algorithms might label variables disproportionally, resulting in a majority class and bias. Decreasing -and preferably eradicating- the important problem has been a major priority, and several research on innovative techniques to overcoming this obstacle have been published. These new techniques can be divided into 4 categories: feature selection, ensemble learning, sampling, and cost-sensitive learning methods.[3]

In biomedical research, web-based applications reliant on ML technology have been increasingly employed for classification and regression tasks. Many of these interactive web-based tools are developed using Shiny, an open-source package in R from RStudio. The primary objective of this research study was to develop a novel user-friendly web tool, based on ML that utilizes a multi-pronged approach to address the class imbalance problem.

MATERIAL AND METHODS

Dataset

The dataset used in this descriptive and retrospective study was obtained from the Cardiovascular Surgery Department of Turgut Özal Medical Center, İnönü University, Malatya, Türkiye. The dataset included records from 929 patients with coronary artery disease (CAD), 149 of which also had hypertension (HT). This research was accepted by the Malatya Clinical Research Ethics Committee, with protocol number of 2016/162 at 8/10/2016. All procedures were performed with respect to the Declaration of Helsinki.

In some studies, when the number of independent variables is 6 or greater in multivariate statistical models, the appropriate sample size can be determined.[4,5]

$n > 104 + k$

condition can be used (n is the number of samples and k is the number of independent variables). The data set used in the study consists of a total of 929 coronary artery patient records, 149 with HT and 780 without HT. Since k=8, it was observed that the above-mentioned condition was met both in the number of patients with and without HT, and in the total number of patients.

A CAD and HT diagnosis require a coronary angiography (at least one coronary stenosis >50% in major epicardial arteries) and a diastolic blood pressure >90 mmHg and/or systolic blood pressure >140 mmHg, respectively.[6] For this study, the classification of HT in patients with CAD was based on eight independent clinical variables, which are described in Table 1.

TABLE 1: Description of clinical variables.

Variables	Variable type	Description	Variable role
Hypertension	Categorical (binary)	Yes/no	Dependent
Age	Numeric	Natural number	Independent
Body surface area	Numeric	Positive real number	Independent
Body mass index	Numeric	Positive real number	Independent
Blood urea nitrogen	Numeric	Positive real number	Independent
Creatinine	Numeric	Positive real number	Independent
Hemoglobin	Numeric	Positive real number	Independent
Hematocrit	Numeric	Positive real number	Independent
Platelet	Numeric	Positive real number	Independent

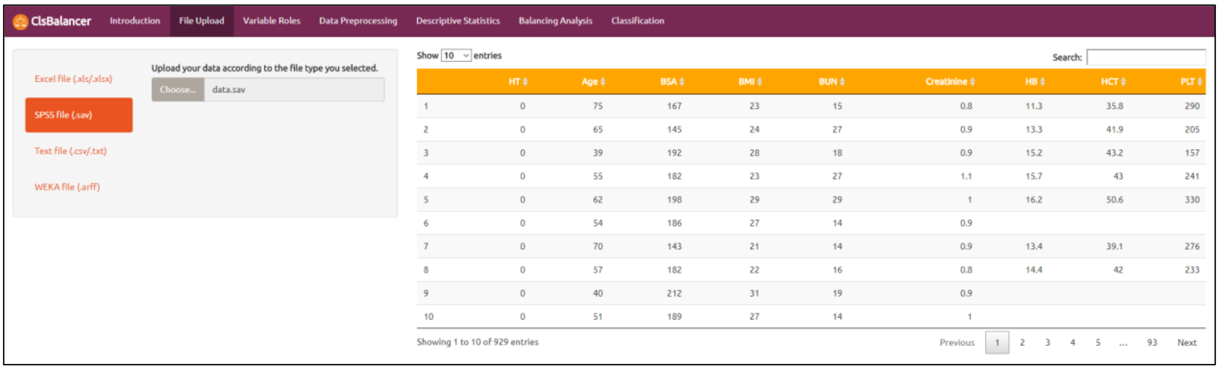
CLASS IMBALANCE PROBLEM

The class imbalance problem is frequently encountered when analyzing datasets in biomedical research and represents one of the most intractable challenges currently facing ML. In this study, a class imbalance problem existed that led to the target label classes not being represented equally.[7] A dataset is regarded as imbalanced if one of the class labels has a small number of observations allotted to it compared to the other label classes. When only two classes are being generated, the minority class contains a smaller number of positive observations and the majority class includes more negative observations.[7,8]

WEB-BASED SOFTWARE

FILE UPLOAD

To utilize this web tool, the dataset file should be uploaded first. The most common file types in data analysis, with various extensions, can be used by the software. Figure 1 contains a screenshot of the "File Upload" menu.



Upload your data according to the file type you selected.

data.sav

Show 10 entries

	HT	Age	BSA	BMI	BUN	Creatinine	HB	HCT	PLT
1	0	75	167	23	15	0.8	11.3	35.8	290
2	0	65	145	24	27	0.9	13.3	41.9	205
3	0	39	192	28	18	0.9	15.2	43.2	157
4	0	55	182	23	27	1.1	15.7	43	241
5	0	62	198	29	29	1	16.2	50.6	330
6	0	54	186	27	14	0.9			
7	0	70	143	21	14	0.9	13.4	39.1	276
8	0	57	182	22	16	0.8	14.4	42	233
9	0	40	212	31	19	0.9			
10	0	51	189	27	14	1			

Showing 1 to 10 of 929 entries

FIGURE 1: The "File Upload" menu.

VARIABLE ROLES

In the "Variable Roles" menu of the related web-based application, the dependent and independent variables of the dataset awaiting analysis are designated.

DATA PREPROCESSING

Next, this web-based tool leads to a "Data Preprocessing" menu with two submenus: "Missing Value Analysis" and "Data Transformation." The first submenu contains 3 approaches for handling missing values: listwise deletion, k-nearest neighbors (k-NN) imputation, and random forest imputation. Through the "Data

Transformation" submenu, the most commonly used data transformation techniques, standardization (z-transformation) and normalization can be applied to the dataset. The "None" option is used when no data transformation is needed.

DESCRIPTIVE STATISTICS

Two submenus are used in the "Descriptive Statistics" menu to relay the results of the statistics can be calculated and tabulated appropriately. In the "ROC-PR Analysis" submenu, the receiver operating characteristic curve and precision-recall curves are graphed; additionally, the associated area under the curve values are generated and tabulated.

BALANCING ANALYSIS

SAMPLING-BASED METHODS

Sampling-based methods are the most commonly used approaches to manage class imbalance problems. The basic idea behind these is to pre-process the training data in a way that minimizes differences in the number of observations between classes. In other words, sampling-based methods change the distribution of the observations in the training dataset to obtain more balance between the minority and majority classes.[3,9] Sampling-based methods are executed using 2 components: oversampling and undersampling. In the oversampling, synthetic data are derived with the help of sampling methods from the minority class observations to obtain a more balanced dataset. On the other hand, in the undersampling, observations from the majority class are removed from the dataset to decrease the class imbalance ratio. Furthermore, oversampling and undersampling approaches can be combined into a hybrid sampling method.

For undersampling methods, various analytical and non-analytical approaches have been proposed in the literatures. The most well-known are random undersampling: Tomek link (TL), condensed nearest neighbor (CNN), one-sided selection (OSS), edited nearest neighbor (ENN), and neighborhood cleaning rule (NCR). The random undersampling approach involves observations from the majority class being randomly extracted from the dataset. Specifically, in the TL approach, 2 observations, x and y, related to the minority and majority classes are taken, respectively. $d(\mathrm{x},\mathrm{y})$ is defined as a metric for the distance between x and y (e.g., Euclidean). If there is no z observation that satisfies the following condition, x and y observations are indicative of a TL.

$$d(\mathrm{x},\mathrm{z}) < d(\mathrm{x},\mathrm{y}) \text{ or } d(\mathrm{y},\mathrm{z}) < d(\mathrm{x},\mathrm{z})$$

If a pair of observations form a TL, those that belong to the majority class can be removed from the dataset.[9,10] The CNN approach aims to extract majority class observations that are far from the decision boundary and, therefore, deemed unnecessary for the training process of the ML model.[11] To achieve this, a coherent subset of the original dataset is created for performing the one-nearest neighbor method. OSS is an undersampling approach that involves sequentially applying the TL and CNN methods.[12] In the ENN method, all observations that are incorrectly classified by a k-NN classifier are extracted, where the k-parameter is usually chosen as k=3.[13,14] The ENN algorithm can extract observations that belong to both the majority and minority classes of the dataset. NCR uses the ENN method mentioned above to determine which of the majority class observations should be removed from the dataset.[15] In the NCR approach, for each observation in the dataset, the three-nearest neighbor observations are determined. If a majority class observation is incorrectly classified by its three-nearest neighbors, it is subtracted from the dataset. If an observation is in a minority category and misclassified by its three-nearest neighbors, afterward the majority label observations between neighbors are removed.

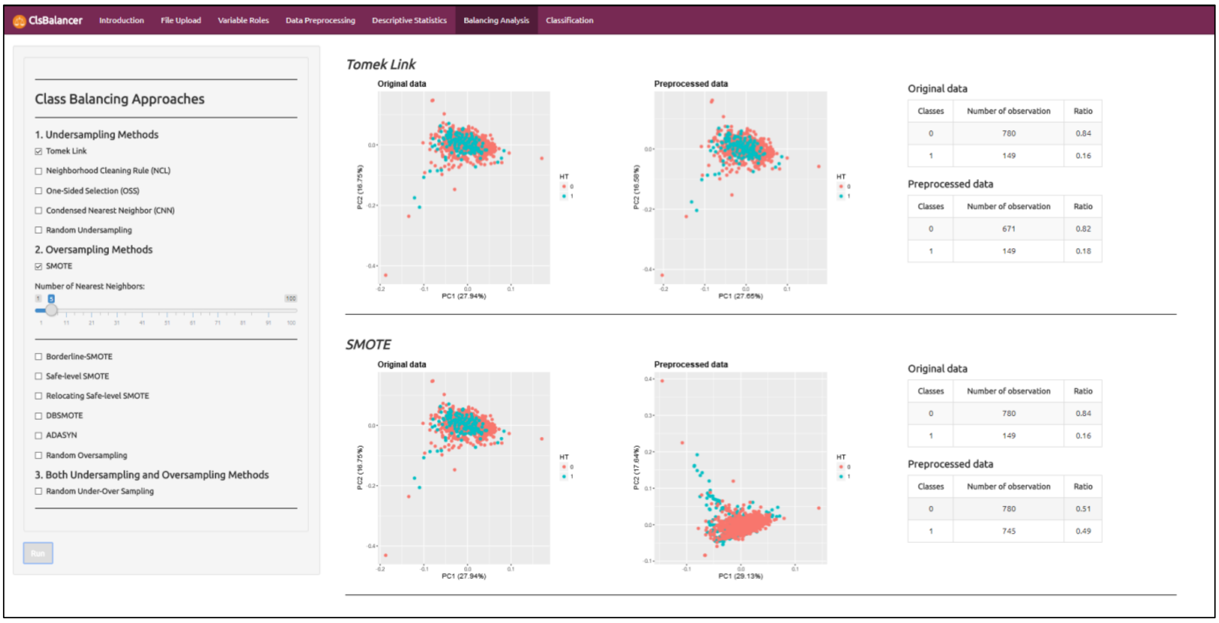
In the undersampling approaches, observations from the minority category are randomly selected and replicated until the number of instances from the minority label equals the number of observations of the majority class. The resulting new dataset thereby contains the same number of observations for both classes.

The synthetic minority over-sampling technique (SMOTE) is a powerful algorithm that has been applied to different disciplines.[16] The SMOTE algorithm produces artificial data on the basis of the similarities of attribute space between available minority observations.[17] To generate the new synthetic minority class observations, SMOTE first randomly selects a minority class observation (a) and detects its minority class k-NN. Then one of the k-NN (b) is randomly selected and a synthetic observation is derived by creating a line segment connecting a to b in the feature space. Synthetic observations are made as a convex combination of 2 selected observations (a and b).[9] The formula for deriving a synthetic observation is as follows:

$$x_{\mathrm{synthetic}} = a + (b - a).\delta \quad (1)$$

Where δ is a randomly selected number in the range [0, 1]. Borderline-SMOTE (BLSMOTE) is an algorithm that directly tries to improve the quality of the results from SMOTE.[18,19] This algorithm relies on the ability to reproduce observations from minority classes in the boundary and near the boundary using the SMOTE algorithm. Safe-level SMOTE (SLSMOTE) assigns a level of confidence to each minority class observation before it derives synthetic observations.[20] While synthetic observations are generated using SMOTE, these observations are located closer to the most secure level (far from minority class observations at the border), ensuring that synthetic observations only occur in safe zones.[19] The density-based SMOTE (DBSMOTE) algorithm is based on a density-based set concept.[21] This algorithm applies the density-based spatial clustering of applications with noise algorithm to the positive (minority) label cluster.[22] The aim of the adaptive synthetic sampling (ADASYN) algorithm is to increase class balance by deriving new observations from the minority class through linear interpolation between available minority class observations.[23] ADASYN is an extension of the SMOTE algorithm and synthetic observations form around the border of the minority and majority classes from within the minority class.[23]

Through the "Balancing Analysis" menu, the dataset with the class imbalance problem can be leveled via different techniques. After the related approach is carried out, the principal component plots of the dataset before and after implementing the manipulation are illustrated. The distributions characteristics of the dependent variable are tabulated in the "Balancing Analysis" menu shown in Figure 2.



Class Balancing Approaches

1. Undersampling Methods

☑ Tomek Link
☐ Neighborhood Cleaning Rule (NCL)
☐ One-Sided Selection (OSS)
☐ Condensed Nearest Neighbor (CNN)
☐ Random Undersampling

2. Oversampling Methods

☑ SMOTE

Number of Nearest Neighbors: 5

☐ Borderline-SMOTE
☐ Safe-level SMOTE
☐ Relocating Safe-level SMOTE
☐ DBSMOTE
☐ ADASYN
☐ Random Oversampling

3. Both Undersampling and Oversampling Methods

☐ Random Under-Over Sampling

Tomek Link

Original data

Classes	Number of observation	Ratio
0	780	0.84
1	149	0.16

Preprocessed data

Classes	Number of observation	Ratio
0	671	0.82
1	149	0.18

SMOTE

Original data

Classes	Number of observation	Ratio
0	780	0.84
1	149	0.16

Preprocessed data

Classes	Number of observation	Ratio
0	780	0.51
1	745	0.49

FIGURE 2: The "Balancing Analysis" menu.

THE CLASSIFICATION

CLASSIFICATION MODELS

The web-based tool's "Classification" primary menu classifies balanced and unbalanced datasets depending on the target variable using three ML models. As classification models, 3 approaches were added to the tool. Multilayer perceptron (MLP), extreme learning machine (ELM), and support vector machines with radial basis function kernel were among them (SVM-RBF). K-fold cross-validation and bootstrap can be utilized as resampling approaches to generate unbiased and optimal results. The performance measures, accuracy, sensitivity, specificity, precision, f-measure and g-mean, were utilized to measure the classification performance of the models.

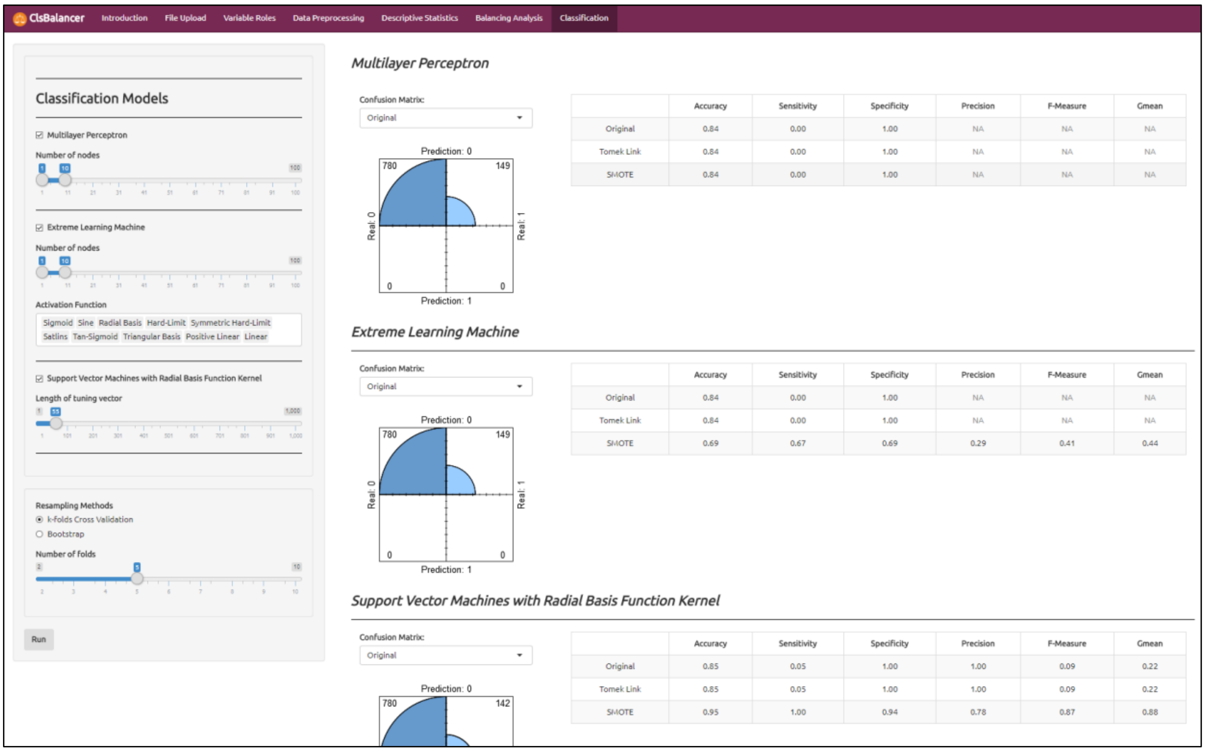
The optimization intervals for the tuning parameters of the classification models used in the software are specified under the relevant models. The optimization parameters for the classification models are tuned using the grid search algorithm. The optimization parameters and the selected optimization intervals for this study are given in Table 2.

The "Balancing Analysis" menu is shown in Figure 3.

TABLE 2: The optimization parameters of the classification models and selected optimization intervals.

Model	**Optimization parameter**	**Optimization intervals**
MLP	□ Number of nodes in hidden layer(s)	□ 1-10
ELM	□ Number of nodes in hidden layer	□ 1-10
	□ Activation function	□ Sigmoid, sine, radial basis, hard-limit, symmetric hard-limit, sigmoid, triangular basis, satlins, positive linear, linear
SVM-RBF	□ Cost parameter (C)	□ 2^{-2}-2^{50}

MLP: Multilayer perceptron; ELM: Extreme learning machine; SVM-RBF: Support vector machines with radial basis function.



Classification Models

☑ Multilayer Perceptron

Number of nodes: 1 – 10 (1–100)

☑ Extreme Learning Machine

Number of nodes: 1 – 10 (1–100)

Activation Function: Sigmoid, Sine, Radial Basis, Hard-Limit, Symmetric Hard-Limit, Satlins, Tan-Sigmoid, Triangular Basis, Positive Linear, Linear

☑ Support Vector Machines with Radial Basis Function Kernel

Length of tuning vector: 55 (1–1,000)

Resampling Methods

● k-folds Cross Validation

○ Bootstrap

Number of folds: 5 (2–10)

Multilayer Perceptron

Confusion Matrix: Original

	Accuracy	Sensitivity	Specificity	Precision	F-Measure	Gmean
Original	0.84	0.00	1.00	NA	NA	NA
Tomek Link	0.84	0.00	1.00	NA	NA	NA
SMOTE	0.84	0.00	1.00	NA	NA	NA

Extreme Learning Machine

Confusion Matrix: Original

	Accuracy	Sensitivity	Specificity	Precision	F-Measure	Gmean
Original	0.84	0.00	1.00	NA	NA	NA
Tomek Link	0.84	0.00	1.00	NA	NA	NA
SMOTE	0.69	0.67	0.69	0.29	0.41	0.44

Support Vector Machines with Radial Basis Function Kernel

Confusion Matrix: Original

	Accuracy	Sensitivity	Specificity	Precision	F-Measure	Gmean
Original	0.85	0.05	1.00	1.00	0.09	0.22
Tomek Link	0.85	0.05	1.00	1.00	0.09	0.22
SMOTE	0.95	1.00	0.94	0.78	0.87	0.88

FIGURE 3: The "Classification" menu.

INTERACTIVE WEB APPLICATION

The interactive web software is freely achievable through http://biostatapps.inonu.edu.tr/twoclsbalancer and this web-based application will be upgrated upon release of R packages, including shiny, shinythemes, shinydashboard, caret, unbalanced, smotefamily, Random Over-Sampling Examples (ROSE), ggplot2, clusterSim, DMwR, clusterSim, precrec, ggfortify, foreign, DataTables (DT), and mice.

RESULTS

The dataset used in the current study consisted of 9 variables: one dependent and 8 independent. The distribution of the dependent variable (HT) was with HT (149, 16.0%) and without HT (780, 84.0%). Also, the imbalance ratio between the HT subclasses was 5.235 (780/149).

Here, the imbalance rate between the 2 classes was obtained by proportioning the number of non-HT patients to that of HT patients. There were a total of 18 missing values in 5 of the independent variables. The random forest algorithm was incorporated into the developed software to assign the imputed values to the observations with missing values.

After eliminating the missing value problem from the dataset, the standardization (z-transform) technique was applied. The imbalance between the classes was then individually eliminated by applying the under-oversampling methods. The number of removed/added observations and the rate of imbalance between newly formed classes after applying TL, Neighborhood Cleaning Rule (NCL), OSS, CNN, random down-sampling, SMOTE, BLSMOTE, SLSMOTE, DBSMOTE, ADASYN, and the random undersampling/oversampling (RUOS) methods are given in Table 3.

TABLE 3: Number of removed/added observations after applying class balancing methods and imbalance ratios between newly formed classes.

Approach	Method	Class distribution of the original dataset (HT, non-HT)		Inter-class imbalance rate of the original dataset	Class distribution after applying the relevant method (HT, non-HT)		Inter-class imbalance rate after applying the relevant method
		Frequency	Frequency		Frequency	Frequency	
Oversampling	TL	780	149	5.235	672	149	4.510
	NCL	780	149	5.235	575	149	3.859
	OSS	780	149	5.235	665	149	4.463
	CNN	780	149	5.235	771	149	5.174
	RUS	780	149	5.235	149	149	1.000
Undersampling	SMOTE	780	149	5.235	780	745	1.047
	BLSMOTE	780	149	5.235	780	772	1.010
	SLSMOTE	780	149	5.235	780	557	1.400
	DBSMOTE	780	149	5.235	780	665	1.173
	ADASYN	780	149	5.235	780	761	1.025
	ROS	780	149	5.235	780	780	1.000
Over+under sampling	RUOS	780	149	5.235	462	467	0.989

HT: Hypertension; TL: Tomek link; NCL: Neighborhood Cleaning Rule; OSS: One-sided selection; CNN: Condensed nearest neighbor; RUS: Random Undersampling; SMOTE: Synthetic minority over-sampling technique; BLSMOTE: Borderline-SMOTE; SLSMOTE: Safe-level SMOTE; DBSMOTE: Density-based SMOTE; ADASYN: Adaptive synthetic sampling; ROS: Random Oversampling; RUOS: Random undersampling/oversampling.

After the balancing analyses, the classification models were applied to the datasets obtained from each approach and to the original dataset (i.e. no class balancing approach applied). The training performance of the relevant classification models was tested using a five-fold cross-validation method. The classification performances of the models are presented in Table 4, including all of the class balancing approaches.

TABLE 4: Classification performance of the three models based on all class balancing approaches.

Model	Approach	Method	Accuracy	Sensitivity	Specificity	Precision	F-measure	G-mean
MLP	None	None	0.84	0	1	-	-	-
	Undersampling	TL	0.84	0	1	-	-	-
		NCL	0.42	0.77	0.35	0.18	0.3	0.38
		OSS	0.84	0	1	-	-	-
		CNN	0.84	0	1	-	-	-
		RUS	0.18	0.99	0.03	0.16	0.28	0.4
	Oversampling	SMOTE	0.82	0.08	0.96	0.29	0.13	0.15
		BLSMOTE	0.84	0	1	-	-	-
		SLSMOTE	0.84	0	1	-	-	-
		DBSMOTE	0.8	0.1	0.94	0.24	0.14	0.15
		ADASYN	0.84	0	1	-	-	-
		ROS	0.83	0.01	0.99	0.15	0.02	0.05
	Over+under sampling	RUOS	0.65	0.52	0.68	0.24	0.32	0.35
ELM	None	None	0.84	0	1	-	-	-
	Undersampling	TL	0.84	0	1	-	-	-
		NCL	0.84	0	1	-	-	-
		OSS	0.84	0	1	-	-	-
		CNN	0.84	0	1	-	-	-
		RUS	0.66	0.64	0.66	0.26	0.37	0.41
	Oversampling	SMOTE	0.68	0.66	0.69	0.29	0.4	0.43
		BLSMOTE	0.68	0.66	0.68	0.28	0.4	0.43
		SLSMOTE	0.75	0.45	0.8	0.3	0.36	0.37
		DBSMOTE	0.71	0.54	0.75	0.29	0.37	0.39
		ADASYN	0.67	0.62	0.68	0.27	0.38	0.41
		ROS	0.7	0.63	0.71	0.29	0.4	0.43
	Over+under sampling	RUOS	0.69	0.62	0.7	0.28	0.39	0.42
SVM-RBF	None	None	0.84	0	1	-	-	-
	Undersampling	TL	0.84	0	1	-	-	-
		NCL	0.84	0	1	-	-	-
		OSS	0.92	0.54	0.99	0.92	0.68	0.71
		CNN	0.84	0	1	-	-	-
		RUS	0.7	1	0.64	0.35	0.52	0.59
	Oversampling	SMOTE	0.99	1	0.99	0.94	0.97	0.97
		BLSMOTE	0.97	0.94	0.98	0.89	0.92	0.92
		SLSMOTE	0.98	0.91	0.99	0.94	0.92	0.92
		DBSMOTE	0.99	0.99	0.99	0.95	0.97	0.97
		ADASYN	0.98	1	0.97	0.88	0.94	0.94
		ROS	0.92	0.87	0.93	0.7	0.78	0.78
	Over+under sampling	RUOS	0.92	0.96	0.92	0.69	0.8	0.81

TL: Tomek link; NCL: Neighborhood Cleaning Rule; OSS: One-sided selection; CNN: Condensed nearest neighbor; RUS: Random Undersampling; SMOTE: Synthetic minority over-sampling technique; BLSMOTE: Borderline-SMOTE; SLSMOTE: Safe-level SMOTE; DBSMOTE: Density-based SMOTE; ADASYN: Adaptive synthetic sampling; ROS: Random Oversampling; RUOS: Random undersampling/oversampling; MLP: Multilayer perceptron; ELM: Extreme learning machine; SVM-RBF: Support vector machines with radial basis function.

As seen in Table 4, the classification performance values from the MLP model on the original dataset were the same as the values from the undersampling methods (TL, OSS, and CNN) and the oversampling approaches (BLSMOTE, SLSMOTE, and ADASYN). After applying the NCL method, the classification performance values of the MLP model were found to be 0.42, 0.77, 0.35, 0.18, 0.30, and 0.38 regarding accuracy, sensitivity, specificity, precision, f-measure, and g-mean, respectively. The results of the SMOTE, DBSMOTE, ADASYN and ROS methods indicated that the performances of the MLP classifications were similar to each other. After applying the RUOS method, the classification performance values of the MLP model were found to be 0.65, 0.52, 0.68, 0.24, 0.32, and 0.35 concerning accuracy, sensitivity, specificity, precision, f-measure, and g-mean, respectively.

The classification performance of the ELM model to the original dataset and the values after TL, NCL, OSS, and CNN were the same as those from the undersampling approaches. The precision, f-measure, and g-mean values could not be calculated because the sensitivity was 0. After applying the RUS approach, the classification performance values from the ELM model were 0.66, 0.64, 0.66, 0.26, 0.37, and 0.41 in terms of accuracy, sensitivity, specificity, precision, f-measure, and g-mean, respectively. When the classification performance of the ELM model was examined after applying the oversampling approaches, the highest accuracy, specificity, and precision values were achieved using SLSMOTE and the highest sensitivity value using BLSMOTE. The highest f-measure and g-mean values were obtained after applying the SMOTE, BLSMOTE, and ROS methods. With the RUOS approach, in which both the oversampling and undersampling methods were randomly implemented, the performance metrics of the ELM model were found to be analogous to the values from the oversampling methods.

As Table 4 details, the classification performance of SVM-RBF model applied to the original dataset and the values after the application of TL, NCL, and CNN were the same as those from the undersampling approaches. The classification metrics for the SVM-RBF model obtained after the OSS approach was applied were calculated to be 0.92, 0.54, 0.99, 0.92, 0.68, and 0.71 for accuracy, sensitivity, specificity, precision, f-measure, and g-mean, respectively. The SVM-RBF classification performance values obtained after RUS were found to be 0.70, 1.00, 0.64, 0.35, 0.52, and 0.59 for accuracy, sensitivity, specificity, precision, f-measure, and g-mean, respectively. After applying the oversampling approaches, the classification performance of SVM-RBF achieved above 0.70 for all performance metrics. With the RUOS approach, the classification metrics of the SVM-RBF model were calculated to be 0.92, 0.96, 0.92, 0.69, 0.80, and 0.81 for accuracy, sensitivity, specificity, precision, f-measure, and g-mean, respectively.

DISCUSSION

The class imbalance question has emerged from the application of ML to many disciplines including bioinformatics, biostatistics, biomedical studies, fraud detection, medical diagnosis, and is regarded as one of the most pressing problems in ML. Essentially, the data including class imbalance issues venture into the learning process, as the classical standard ML techniques assume equal class distribution and nearly equal misclassification cost.[24] The research presented here intended to develop a novel web-based program that can perform many sampling-based methods (i.e. oversampling, undersampling, and hybrid sampling) toward alleviating the class imbalance problem. Additionally, instead of applying a single classifier, this novel web-based tool includes three ML models (SVM-RBF, ELM, and MLP) that evaluate both the classification performances of the models and the efficacy of the data balancing methods. In this regard, our research more extensively addresses the imbalance class problem in ML-based classification than other studies to date.[25,26] Our web-based software's ability to implement advanced, complex techniques in ML toward selecting the best model(s) for classification demonstrate the originality and validity of this research.

All we know, this is the first research to classify HT in patients with CAD and develop a web-based, interactive solution for tackling the class imbalance problem. The dataset used in this research suffered a class imbalance problem in terms of the distribution of the HT dependent variable (HT group: 16%

compared to the non-HT group: 84%) in patients with CAD. This important issue has been widely encountered in biomedical research in recent years. This problem has likely caused the classification algorithms to generate biased results that are easily misinterpreted by researchers. Therefore, the current research attempted to mitigate the problem of class imbalance using sampling-based class balancing methods (i.e. oversampling, undersampling, and hybrid sampling). Toward this aim, user-friendly software was developed to aid in avoiding this pervasive problem.

To test the functionality of the web-based application developed within the scope of this study, we applied the sampling-based methods to a clinical dataset plagued by a class imbalance problem with an inter-class imbalance ratio of 5.235:1. After applying the sampling-based methods, the classification process was implemented using the MLP, ELM, and SVM-RBF algorithms. When the classification models were applied to the original dataset with class imbalance, the accuracy, sensitivity, and specificity values were the same for all three models. Because the sensitivity value was 0.00, the precision, f-measure, and g-mean could not be calculated. When the accuracy scores of the 3 models (0.84) were taken into consideration, it became apparent that the models successfully generated unbiased classification results, except for the accuracy measure, under the influence of the majority class, as the number of patients in the non-HT group was greater than the number of patients in the HT group. Therefore, the constructed models gave biased and misleading results in each model.

When the undersampling methods were applied and the resulting classification performance of the 3 models compared, it was evident that the majority of the class observations extracted from the dataset by the TL and CNN methods did not enhance classification ability as the results were the same as when no correction methods were used. A similar situation was observed for the NCL method in the ELM and SVM-RBF models and the OSS method in the MLP and ELM models. Applying the NCL method in the MLP model confirmed that the classification performance values were not at the desired level, except for the sensitivity value (0.77). Applying the RUS method had an adverse impact on the prediction performance of the MLP model and lowered the other performance metrics while raising the sensitivity value.

There are several biomedical publications on the classification of HT using multiple ML techniques.[27-32] Nevertheless, few researchers have reported classifying HT data affected by a class imbalance problem. A study developed and validated a risk prediction method for HT and has constructed an ensemble of the classification of trees.[28] The above-mentioned study concludes that their analytics model enhances HT care. An additional study developed a novel technique, called BrSmoteSvm, designed for HT classification with class imbalance problems.[25] Their research indicated that the BrSmoteSvm method outperformed other multi-class classifiers in several performance metrics, one of which was a lower precision value (0.66 compared to 0.95 for their newly generated model) calculated from the present study.

Some methods were not included in the current study (e.g., feature selection-based, ensemble learning-based, cost-sensitive learning-based methods, and so forth); these will be examined in future studies. Specifically, recent state-of-the-art ML algorithms [e.g. DeepBoost (Google Inc., USA) XGBoost (Distributed (Deep) Machine Learning Community (DMLC) group, USA)] coupled with sampling techniques will be studied with the ensemble models.[33,34] Then, the above-mentioned techniques will be integrated into the web-based tool developed here and made publicly available through subsequent updates. Thus, the software will continue to produce more robust and consistent classification results.

CONCLUSION

Based on the results of this study, oversampling methods enabled successful classification performance more than undersampling methods. When the undersampling methods were applied, the three classification models could not accurately classify the data, whereas when the oversampling methods were implemented, the SVM-RBF model outperformed the other 2 models. Taken together, the web-based application proposed, developed, and validated in this study represents a promising tool for medical researchers and clinicians alike with the potential to greatly facilitate biotechnology-based efforts toward improving human health and well-being.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

***Idea/Concept:** Ahmet Kadir Arslan, Cemil Çolak, Mehmet Cengiz Çolak; **Design:** Ahmet Kadir Arslan, Cemil Çolak, Mehmet Cengiz Çolak; **Control/Supervision:** Ahmet Kadir Arslan, Cemil Çolak, Mehmet Cengiz Çolak; **Data Collection and/or Processing:** Ahmet Kadir Arslan, Cemil Çolak, Mehmet Cengiz Çolak; **Analysis and/or Interpretation:** Ahmet Kadir Arslan, Cemil Çolak, Mehmet Cengiz Çolak; **Literature Review:** Ahmet Kadir Arslan, Cemil Çolak, Mehmet Cengiz Çolak; **Writing the Article:** Ahmet Kadir Arslan, Cemil Çolak, Mehmet Cengiz Çolak; **Critical Review:** Ahmet Kadir Arslan, Cemil Çolak, Mehmet Cengiz Çolak; **References and Fundings:** Ahmet Kadir Arslan, Cemil Çolak, Mehmet Cengiz Çolak; **Materials:** Mehmet Cengiz Çolak.*

REFERENCES

1. Sagiroglu S, Sinanc D. Big data: A review. International Conference on Collaboration Technologies and Systems (CTS). 2013;42-7. [Crossref]
2. Firat F, Arslan AK, Colak C, Harputluoglu H. Estimation of risk factors associated with colorectal cancer: an application of knowledge discovery in databases. Kuwait J. Sci. 2016;43(2):151-61. [Link]
3. Bekkar M, Alitouche TA. Imbalanced data learning approaches review. Int J Data Min Knowl Manag Process. 2013;3(4):15-33. [Crossref]
4. Alpar CR. Uygulamalı Çok Değişkenli İstatistiksel Yöntemler. 4. Baskı. Ankara: Detay Yayıncılık; 2013.
5. Sümbüloğlu V, Sümbüloğlu K. Klinik Saha Araştırmalarında Örnekleme Yöntemleri ve Örneklem Büyüklüğü. 1. Baskı. Ankara: Hatiboğlu Yayınevi; 2005.
6. Colak MC, Colak C, Kocatürk H, Sağiroğlu S, Barutçu I. Predicting coronary artery disease using different artificial neural network models. Anadolu Kardiyol Derg. 2008;8(4):249-54. [PubMed]
7. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-Level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: Theeramunkong T, Kijsirikul B, Cercone N, Ho TB, eds. Advances in Knowledge Discovery and Data Mining. 1st ed. Thailand: Springer Berlin Heidelberg; 2009. p.475-82. [Crossref]
8. Cohen G, Hilario M, Sax H, Hugonnet S, Geissbuhler A. Learning from imbalanced data in surveillance of nosocomial infection. Artif Intell Med. 2006;37(1):7-18. [Crossref] [PubMed]
9. He H, Ma Y. Imbalanced Learning: Foundations, Algorithms, and Applications. 1st ed. USA: John Wiley & Sons; 2013. [Crossref]
10. Tomek I. An experiment with the edited nearest-neighbor rule. IEEE Trans syst Man Cybern. 1976;6(6):448-52. [Crossref]
11. Hart P. The condensed nearest neighbor rule. IEEE Trans Inf Theory. 1968;14(3):515-6. [Crossref]
12. Kubat M, Matwin S. Addressing the curse of imbalanced training sets: one-sided selection. In: Fisher DH, ed. Proceedings of the 14th International Conference on Machine Learning. USA: Morgan Kaufmann Publishers Inc.; 1997. p.179-86. [Link]
13. Wilson DL. Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans syst Man Cybern. 1972;2(3):408-21. [Crossref]
14. García-Borroto M, Villuendas-Rey Y, Carrasco-Ochoa JA, Martínez-Trinidad JF. Using maximum similarity graphs to edit nearest neighbor classifiers. In: Corrochano EB, Eklundh JO, eds. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. 1st ed. Springer; 2009. p.489-96. [Crossref]
15. Laurikkala J. Improving identification of difficult small classes by balancing class distribution. In: Quaglini S, Barahona P, Andreassen S, eds. Artificial Intelligence in Medicine. 1st ed. Portugal: Springer-Verlag; 2001. p.63-6. [Crossref]
16. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16(1):321-57. [Crossref]
17. He H, Garcia EA. Learning from imbalanced data. IEEE Trans knowl data eng. 2009;21(9):1263-84. [Crossref]
18. Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang DS, Zhang XP, Huang GB, eds. 1st ed. Advances in Intelligent Computing. China: Springer; 2005. p.878-87. [Crossref]
19. Verbiest N, Ramentol E, Cornelis C, Herrera F. Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection. Appl Soft Comput. 2014;22(2):511-7. [Crossref]
20. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: Theeramunkong T, Kijsirikul B, Cercone N, Ho TB, eds. Advances in Knowledge Discovery and Data Mining. 1st ed. Berlin, Heidelberg: Springer Berlin, Heidelberg; 2009. p.475-82. [Crossref]
21. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. DBSMOTE: Density-Based Synthetic Minority Over-sampling TEchnique. Appl Intell. 2012;36(3):664-84. [Crossref]
22. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland, Oregon, USA: 1996. p.226-31. [Link]
23. He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. IEEE World Congress on Computational Intelligence. 2008;1322-8. [Link]
24. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res. 2017;18(1):559-63. [Link]

25. Li GZ, He Z, Shao FF, Ou AH, Lin XZ. Patient classification of hypertension in Traditional Chinese Medicine using multi-label learning techniques. BMC Med Genomics. 2015;8 Suppl 3(Suppl 3):S4. [Crossref] [PubMed] [PMC]
26. Li GZ, Yan SX, You M, Sun S, Ou A. Intelligent ZHENG Classification of Hypertension Depending on ML-kNN and Information Fusion. Evid Based Complement Alternat Med. 2012;2012:837245. [Crossref] [PubMed] [PMC]
27. Antalek MD, Suwa K, Schaffer M, Fenster B, Markl M, Freed B, et al. Non-invasive classification of pulmonary hypertension using 4D flow MRI and random forests. Circulation. 2017;136(1). [Link]
28. Ye C, Fu T, Hao S, Zhang Y, Wang O, Jin B, et al. Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. J Med Internet Res. 2018;20(1):e22. [Crossref] [PubMed] [PMC]
29. LaFreniere D, Zulkernine F, Barber D, Martin K. Using machine learning to predict hypertension from a clinical dataset. IEEE Symposium Series on Computational Intelligence (SSCI). 2016;1-7. [Crossref]
30. Kublanov VS, Dolganov AY, Belo D, Gamboa H. Comparison of machine learning methods for the arterial hypertension diagnostics. Appl Bionics Biomech. 2017;2017:5985479. [Crossref] [PubMed] [PMC]
31. Seffens W, Evans C; Minority Health-GRID Network, Taylor H. Machine learning data imputation and classification in a multicohort hypertension clinical study. Bioinform Biol Insights. 2016;9(Suppl 3):43-54. [Crossref] [PubMed] [PMC]
32. Held E, Cape J, Tintle N. Comparing machine learning and logistic regression methods for predicting hypertension using a combination of gene expression and next-generation sequencing data. BMC Proc. 2016;10(Suppl 7):141-5. [Crossref] [PubMed] [PMC]
33. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. arXiv. 2016:785-94. [Crossref]
34. Cortes C, Mohri M, Syed U. Deep boosting. PMLR. 2014;32(2):1179-87. [Link]