Kafkas Universitesi Veteriner Fakultesi Dergisi
Journal Home-Page: http://vetdergikafkas.org
E-ISSN: 1309-2251

Kafkas Univ Vet Fak Derg
30 (1): 31-39, 2024
DOI: 10.9775/kvfd.2023.30325

Research Article

Comparison of Some Balancing Methods for Classification of Pacing Horses Using Tree-based Machine Learning Algorithms

Hülya ÖZEN [1 (*)] Doğukan ÖZEN [2] Banu YÜCEER ÖZKUL [3] Ceyhan ÖZBEYAZ [3]

[1] University of Health Sciences, Gulhane Faculty of Medicine, Department of Medical Informatics, TR-06018 Ankara - TÜRKİYE
[2] Ankara University, Faculty of Veterinary Medicine, Department of Biostatistics, TR-06070 Ankara - TÜRKİYE
[3] Ankara University, Faculty of Veterinary Medicine, Department of Animal Science, TR-06070 Ankara - TÜRKİYE

(*) **Corresponding author:** Hülya ÖZEN
Phone: +90 312 567 1500-4037
E-mail: hulya.ozen@sbu.edu.tr

How to cite this article?
Özen H, Özen D, Yüceer Özkul B, Özbeyaz C: Comparison of some balancing methods for classification of pacing horses using tree-based machine learning algorithms. *Kafkas Univ Vet Fak Derg,* 30 (1): 31-39, 2024.
DOI: 10.9775/kvfd.2023.30325

Article ID: KVFD-2023-30325
Received: 27.07.2023
Accepted: 30.10.2023
Published Online: 09.12.2023

ABSTRACT

Classifiers in machine learning work on the principle that the observations are evenly distributed across the classes. However, real-world datasets frequently exhibit skewed distributions of classes, which is called imbalanced, causing the classifiers make highly biased predictions. One of the several method groups that deal with imbalance data problem is class balancing methods. We aimed to compare some class balancing methods during the classification of pacing horses according to their origins. Data set contains morphological traits of horses and four origin classes with different sample sizes that leads a multi-class imbalanced data problem. Training data set was modified with different balancing methods. Each balanced data set was trained with C5.0, Random Forest and Extreme Gradient Boosting Machine classifiers. Method comparisons were made based on comparison metrics using the original test set. The best prediction result was obtained on the data set balanced with random undersampling method regarding both G-mean and Matthews Correlation Coefficient; however, the best result according to F1 score was observed on the data set balanced with Adaptive Synthetic Sampling Approach (ADASYN). Primary important variables of the best models were body length, withers height, chest circumference and rump height. The Bulgarian origin was the most accurately predicted class despite having the smallest sample size. Class balancing methods clearly improved the performance of classifiers for predicting origins of pacing horses.

Keywords: Class balancing methods, Imbalanced data, Machine learning, Multi-class classification, Pacing horses

Introduction

Classification is a supervised learning technique in machine learning, where the model attempts to predict a proper label or class for a given data. Recently, classification algorithms or classifiers have been employed in veterinary medicine [1-4], as well as in areas such as disease identification [5,6] and fraud detection [7].

Many classifiers provide high prediction performance when they work with a balanced data set, where the numbers of observations are almost equal in each class. However, real world data sets are commonly *imbalanced*, in which one class could be represented by a lot of observations called *majority class*, while the others are only represented by a few called *minority class* [8]. Many studies proposed some rules such as *imbalance ratio* to define a data set as *imbalanced*. The imbalanced ratio (IR) is defined as the ratio of number of observations of the majority class to minority class [9]. Although several IR values have been proposed, there is still not a clear rule [10]. For instance, in Fernandez et al.'s study [11], datasets with an IR greater than 1.5 are regarded as imbalanced.

According to several studies, classifiers can perform well on imbalanced data sets with clear class separation, since the main idea behind classifiers is to find optimum decision boundaries [12,13]. The actual issue that causes classifiers to struggle with imbalanced data sets reveals as overlapping regions. Overlapping regions occur when data points from various classes are relatively near to each other or when class boundaries overlap. These regions have an adverse impact on the classification task, since they reduce the representative power of the minority classes with small sample sizes [13,14].

In order to effectively deal with multiclass imbalance problem, some approaches are proposed and grouped

under three main titles, (i) data preprocessing methods, (ii) inbuilt mechanisms and (iii) cost sensitive methods. Data preprocessing methods also named as class balancing methods involve modifying the original dataset to create a more balanced class distribution [11,13].

Pacing horses has a popular place in Türkiye. While they were once used for transportation, they now mostly compete in races [15,16]. Despite the fact that Turkish Native breeds are frequently seen in the field, new origins including Iranian, Afghan, and Bulgarian horses have lately been introduced and have found a place among the other pacing horses. Apart from their pacing, the common or distinguishing traits of these horses are still limited and were only subjected to morphological comparison with classical approaches by Yüceer et al.[16] and Çağlayan et al.[15] so far. Therefore, identifying the differences and classifying these origins correctly created a new field of study. Accurate classification according to origin is essential for breed preservation, breed management and genetic improvement. It may also contribute to the cultural heritage and traditions associated with these horses.

This paper attempted to address the classification challenges associated with the origin of pacing horses using morphological traits, and the effectiveness of different balancing methods in improving classification accuracy. In this context, the purpose of this study was (i) to compare and evaluate some class balancing methods during classifying the origins of pacing horses, (ii) to examine the predictive performance of tree-based classifiers, (iii) to assess the relative variable importance values of the best predictive classifiers, and (iv) to draw the attention of those who encounter the problem of class imbalance in the field of veterinary medicine and to offer a solution.

Material and Methods

Data Set

The data set used in this study consists of 430 pacing horses raised in different geographical regions of Türkiye and aged 4 years or older. The class variable is the origin of pacing horses, namely Iranian, Afghan, Bulgarian and Turkish Native. Classifiers are trained on morphological traits such as body length, cannon bone circumference, chest circumference, chest depth, head length, rump height and withers height. Detailed summary statistics are given in *Table 1*. Turkish Native origin that has the highest number of observations is the majority class, while the others are minority classes. A multi-class imbalanced situation is indicated by the IR values, which are 4.79 for Iranian, 10.53 for Afghan, and 17.56 for Bulgarian origin.

Class Balancing Methods

Class balancing methods have the advantage of being more adaptable because their use is independent of the classifier chosen [11]. Many class balancing methods are proposed in previous studies that concentrate on modifying the training data to build an effective classifier. In terms of balancing data sets, we can differentiate between methods that create new observations for minority classes are called oversampling and those that eliminate observations from the majority class are called undersampling. Some combinations of these methods are also commonly used.

Table 1. *Summary statistics of morphological traits [Mean ± Standard deviation; Median (Minimum-Maximum)]*

Morphological Traits	Origin			
	Iranian (n=66)	**Afghan (n=30)**	**Bulgarian (n=18)**	**Turkish Native (n=316)**
Body length	150.25±4.44	148.33±4.6	165.78±5.98	145.49±5.78
	150 (141-163)	148 (142-160)	167 (156-173)	145 (131-164)
Cannon bone circumference	17.63±0.68	17.57±0.57	19.28±0.73	17.07±0.88
	17.5 (15.5-19.5)	17.5 (16.5-18.5)	19 (18.5-21)	17 (14-20)
Chest circumference	159.12±4.64	157.47±4.1	179.11±9.34	156.44±7.15
	159 (148-175)	156.5 (153-167)	180.5 (162-200)	156 (137-174)
Chest depth	63.83±2.11	63.6±2.71	70.44±2.45	61.92±3.31
	64 (58-68)	63.5 (58-68)	71 (65-75)	62 (51-77)
Head length	53.41±1.7	53.05±1.32	57.28±2.02	52.54±1.63
	53 (50-57)	53 (51-55)	57 (53-61)	53 (47-57)
Rump height	144.24±3.43	142.68±3.44	157.56±4.05	139.67±4.67
	144 (136-153)	142 (138-152)	159 (151-163)	140 (127-151)
Withers height	143.65±3.14	142.97±2.52	156.39±3.76	138.92±4.87
	143 (135-151)	143 (139-149)	156 (147-162)	139 (123-151)

In this paper, nine different class balancing methods were used, which are determined based on their superior results in previous studies [12,17]. Random oversampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic Sampling Approach (ADASYN) were oversampling methods while Random undersampling (RUS), Tomek links (TL), One-sided selection (OSS), and Edited nearest neighbor (ENN) were used as undersampling methods. SMOTE+TL and SMOTE+ENN were used as combination methods. Although balancing methods are initially developed for binary classification, their application has expanded to multiclass scenarios with pairwise class implementations. Methods are briefly described below.

ROS: This is a non-heuristic method that replicates minority class observations randomly to balance the classes. Though it is easy to implement, it may increase the overfitting [11,14].

SMOTE: It is an oversampling method that contributes new observations to the minority classes without replicating. SMOTE uses interpolation technique that is a type of estimation, where new data points are created within the range of known data points [18]. As a result, with using SMOTE the overfitting issue is avoided. However, it could result in the minority class's decision boundaries expanding into the space of the majority class [12].

ADASYN: Unlike the SMOTE method, it gives weight to minority class observations that are difficult to classify and uses less of those that can be classified easily. It performs the interpolation with the minority class observations and the nearest minority or majority class neighbor observations [19].

RUS: This is also a non-heuristic method that removes majority class observations randomly to balance the classes. Disadvantage of this method is that it can eliminate potentially useful observations and leads to information loss [11,12,14].

TL: Let two observations be e_i and e_j belonging to different classes with the distance $d(e_i, e_j)$ between them. A pair (e_i, e_j) is called TL, if there is no example e_l, such that $d(e_i,e_l) < d(e_i, e_j)$ or $d(e_j,e_l) < d(e_i, e_j)$. TL can be applied as an undersampling method or as a data cleaning method. As an undersampling method, only observations of the majority class are eliminated, and as a data cleaning method, examples of both classes are removed [20].

OSS: This is a two-stage undersampling method. After the results obtained from the application of TL, Condensed Nearest Neighbor (CNN) rule is applied on the observations. TL is used as undersampling method to remove the noisy and borderline observations of majority class. CNN aims to remove examples from the majority class that are distant from the decision border [21].

ENN: It is an undersampling method that uses k nearest neighbor method, where k is equal to 3. This method eliminates a majority class observation unless there are more majority class observations among its three nearest neighbors [22].

SMOTE+TL: Creating synthetic observations with SMOTE can make minority class observations to expand too close to the majority class space. To create better-defined class observations, TL is applied to the over-sampled data set as a data cleaning method. Thus, not only majority class observations, but also minority class observations are removed [12].

SMOTE+ENN: The idea of this method is similar with SMOTE+TL. First SMOTE is applied on data set to create synthetic observations for minority class. Then ENN is applied as a data cleaning method by removing observations from both majority and minority classes [12].

Classification Methods Used in the Study

Tree-based algorithms were used in this study. C5.0 was preferred as single tree. Extreme Gradient Boosting Machine (XGBM) and Random Forest (RF) were regarded as tree-based ensemble learning methods, where XGBM belongs to the boosting family, and RF to the bagging family. The remaining part of this section provides a brief description of the classifiers chosen for our study and how they are applied to the dataset.

C5.0 is s single tree that is an extent work of C4.5 decision tree [23]. It provides high accuracy by using boosting technique. C5.0 has strong opinions about pruning and handles a lot of the choices automatically using defaults that are generally acceptable.

RF is a commonly used ensemble learning method that combines various decision trees to produce a single outcome. Diversity of the trees in the RF is based on two characteristics: bootstrapping the original training data and selection of a random subset of the variables at each split during tree building. Final outcome of the RF is decided based on averaging or majority voting of the trees for regression and classification, respectively [24,25].

XGBM is a more accurate, fast and scalable implementation of gradient boosting decision trees (GBDT) that train an ensemble of decision trees iteratively with boosting technique. The concept behind gradient boosting is to use gradient descent algorithm over an object function to combine a single weak classifier with other weak classifiers to build a strong classifier. In this process, it is aimed to minimize the prediction error considerably [26].

The existence of multiple classes implies an extra challenge for machine learning algorithms because the boundaries of the classes may overlap that leads poor performance.

Class binarization techniques are used to convert the initial multiple-class problem into binary subsets that are simpler to distinguish. In this study, multi-class issue is divided into more straightforward binary classification tasks with one-versus-all approach. Solutions are created to deal with two-class imbalanced datasets for each binary classification task [11].

In this study, original data set is randomly divided into the training and test set with 70% and 30 % ratio, respectively. Training set is used to train the models, while test set is used for testing the examined model for performance comparison purposes. Before building the models, class balancing methods are applied on the original training set. Test set is retained original, while different balanced training sets are created. Although there are many settings that can be used with each machine learning algorithm, we choose the best configuration based on parameter tuning, which offers the parameter set with the best prediction on train sets. Parameter tunning is carried out with applying 10 times repeated 10- fold cross validation technique on the training set. In each repeat, 10-fold cross-validation is applied by splitting the training data 10 equal folds. Each fold is used as validation set for the trained model where the remaining folds are used as training. In addition to determining the best class balancing method, relative variable importance values of the classifiers offering the highest performance are also given. These values present which variables significantly impact the model performance and its predictions [25,27,28].

All calculations in this study are performed with R version 4.2.2 [29] using R Studio (version: 2022.07.2+576) [30]. The R packages Caret (Classification and Regression Training) [28] and UBL (Utility-Based Learning) [31] are used for training the models for each examined classifier and balancing training data sets, respectively.

Model Comparison Metrics

Although various metrics have been proposed over time to compare the performance of classifiers, not all of them are suitable for use in the case of class imbalance. Some metrics such as accuracy or error rate are biased in favor of majority classes [9,32]. The most frequently used metrics in imbalanced data problems are precision, recall, F1 score, G-mean, and Matthews Correlation Coefficient (MCC) [13,32].

The comparison metrics are computed from a confusion matrix that has two dimensions. One dimension is represented by the actual class of an observation and the other by the class that the classifier predicts. The metrics for binary classification are calculated for each class with using a 2x2 confusion matrix. As a result, names of the classes are changed to *positive* and *negative* class, with positive class denoting the class of interest and negative class denoting the other class. Precision measures the correctly classified positive class. Recall presents the proportion of correctly predicted of all actual positive observations. F1 score is a harmonic mean of precision and recall [33]. Calculation of F1 score is given in (1).

$$F1\ score = \frac{2*Precision*\ Recall}{Precison+Recall} \qquad (1)$$

Precision, recall and F1 score can be expanded to multi-class scenarios by using micro-averaging or macro-averaging methods, which provides a single output for all classes [32]. In this study, macro averaging method was used by weighting the metrics with the number of observations in the classes, that was suggested for multi-class imbalanced scenarios [34].

G-mean is a comparison metric that is calculated by taking geometric mean of the recall values of the classes. An approach introduced by Sun et al.[35] for multi-class scenarios is used in this study. The formula is presented in (2), where c denotes the number of classes.

$$G-mean = \left(\prod_{i=1}^{C} Recall_i\right)^{\frac{1}{C}} \qquad (2)$$

MCC is proposed by Halimu et al.[36] for binary classification problems. It provides a correlation coefficient between actual and predicted observations. MCC also can be expanded to multi-class scenario as MMCC with combining pairwise MCC values of the classes, which is given in (3).

$$MMCC = \frac{2}{C(C-1)} \sum_{i<j} MCC_{(i,j)} \qquad (3)$$

When the calculations of the metrics are examined, if the results of all observations are collected in the negative class, it is possible for F1 score and MMCC to get results that go to infinity [32].

Results

Number of observations for each class in the original and balancing methods applied training sets were given in *Table 2*. Applying oversampling methods made a great increase in the numbers of observations of minority classes. With the use of ROS or RUS, all classes had the same number of observations. ENN made a greater decrease in the number of observations of Turkish Native class with respect to TL and OSS. The use of combined methods resulted in an increase in the number of observations in minority classes and a decrease in the number of observations in the majority class Turkish Native.

The performance results of the classifiers for each balancing method were given in *Table 3*. A non-computable issue was indicated by the label NaN. This situation was met when all observations are collected in one class.

The results of the original data set, where the classes are highly skewed, indicated that C5.0 was the best classifier

Table 2. *Number of observations for each class in the original and balanced training sets*

Category	Balancing Method	Origin			
		Iranian	Afghan	Bulgarian	Turkish Native
Original		47	21	13	222
Oversampling Methods	ROS	222	222	222	222
	ADASYN	226	216	222	222
	SMOTE	211	220	221	222
Undersampling methods	RUS	13	13	13	13
	TL	47	21	13	202
	OSS	47	21	13	197
	ENN	47	21	13	185
Combined methods	SMOTE+TL	205	219	221	215
	SMOTE+ENN	195	208	220	146

among the others with the highest F1 score, G-mean and MMCC values. When the results of the oversampling methods were examined, highest F1 score and MMCC values were observed on the data set balanced with ADASYN, where SMOTE provided the highest G-mean. Classifiers yielded the lowest metric results on the data set balanced with ROS among the oversampling methods. XGBM was the most successful classifier on the data sets balanced with ROS and ADASYN, while RF performed best on the dataset balanced with SMOTE.

As the results of undersampling methods were evaluated, the highest G-mean and MMCC values were obtained on the dataset balanced with RUS, while the highest F1 score was provided on the data set balanced with OSS. C5.0 demonstrated the best performance on the data sets balanced with TL, OSS, and ENN, where tree-based ensemble learning methods were failed.

In consideration with the combined balancing methods, Classifiers outperformed on the data set balanced with SMOTE+ENN than the data set balanced with SMOTE+TL, according to the results of G-mean and MMCC. On the contrary, F1 score indicated that SMOTE+TL was the better combined method.

When the whole class balancing methods were evaluated, the highest values of G-mean and MMCC were observed on the data set balanced with RUS and trained with RF and C5.0, respectively. However according the F1 score, the best performance was observed on the data set balanced with ADASYN and trained using XGBM. Furthermore, the lowest value of F1 score were observed on the data set balanced with RUS. Unlike F1 score, the lowest values of G-mean and MMCC were obtained on the data set balanced with ROS. The precision values of all classes varied between 0.66863 and 0.796, whereas the recall values varied between 0.56696 and 0.80311 *(Table 3)*. It was demonstrated that F1 score and MMCC were directly affected with zero values in the confusion matrix, took values between 0.63872 and 0.76816, and 0.42429 and 0.63915, respectively.

Variable importance values were presented in *Fig. 1* to assess the contribution of each morphological trait in the best predictive classifiers according to the highest values of G-mean, MMCC and F1 score, respectively. Body length, withers height, chest circumference and rump height were seemed to be the common primary used predictors during the training classifiers.

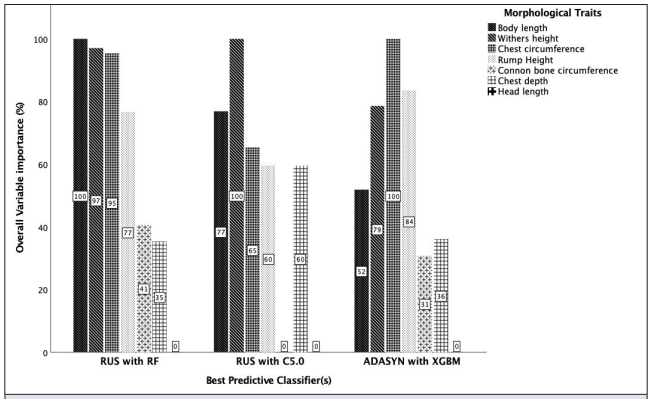

Fig 1. Variable importance values of the best predictive classifiers according to G-mean, MMCC and F1 score

Fig. 2 shows the distributions of the classes corresponding to the original data and the data sets balanced with RUS and ADASYN using the two most significant predictors provided in *Fig 1*. It was found that Turkish Native horses had lower values whereas Bulgarian horses had greater values. Iranian and Afghan horses fell between aforementioned classes. Values of Turkish Native horses seemed to be overlapping on the values of Iranian and Afghan horses on the original training set *(Fig. 2-a)*. As the three graphs were compared, it was seen that the influence of the Turkish Native class on the Iranian and Afghan classes had decreased in *Fig. 2-b*, on which a simpler

Table 3. *Comparison metric results of models for original data and data sets with balancing methods applied*

Category	Balancing Method		Precision	Recall	F1 Score	G-Mean	MMCC
Original		C5.0	0.70322	0.72438	0.70583	0.41825	0.56523
		RF	0.68360	0.65112	NaN	0	NaN
		XBGM	NaN	0.78741	NaN	0	NaN
Oversampling Methods	ROS	C5.0	0.69634	0.66139	0.67738	0.36835	0.42429
		RF	0.67892	0.69289	0.68376	0.34845	0.47818
		XBGM	0.72840	0.66143	0.68970	0.43202	0.54117
	ADASYN	C5.0	0.68321	0.67716	0.67918	0.41039	0.51255
		RF	0.72275	0.74017	NaN	0	NaN
		XBGM	0.76072	0.77951	**0.76816**	0.46568	0.63222
	SMOTE	C5.0	0.74257	0.66143	0.69380	0.49690	0.52644
		RF	0.73781	0.68504	0.70706	0.50214	0.53024
		XBGM	0.66863	0.66928	NaN	0.00000	NaN
Undersampling methods	RUS	C5.0	0.79369	0.56696	0.63872	0.60548	**0.63915**
		RF	0.78624	0.66143	0.70237	**0.65315**	0.63306
		XBGM	0.79600	0.61417	0.67712	0.57847	0.57826
	TL	C5.0	0.72254	0.74799	0.72932	0.46557	0.61450
		RF	0.68316	0.74804	NaN	0	NaN
		XBGM	NaN	0.80311	NaN	0	NaN
	OSS	C5.0	0.73611	0.74802	0.73658	0.49880	0.63213
		RF	0.69704	0.74802	NaN	0	NaN
		XBGM	NaN	0.79527	NaN	0	NaN
	ENN	C5.0	0.72802	0.70079	0.71304	0.50543	0.60556
		RF	0.72542	0.73774	NaN	0	NaN
		XBGM	0.72860	0.77164	NaN	0	NaN
Combined methods	SMOTE+TL	C5.0	0.71800	0.66931	0.68935	0.42070	0.46455
		RF	0.73637	0.70076	0.71647	0.45115	0.56610
		XBGM	0.69039	0.70077	NaN	0	NaN
	SMOTE+ENN	C5.0	0.77362	0.62204	0.67517	0.44493	0.46114
		RF	0.78699	0.64564	0.69118	0.59955	0.63179
		XBGM	0.73502	0.68503	0.70475	0.45843	0.58852

NaN: *Not a Number;* ***ROS:*** *Random Over-Sampling;* ***ADASYN:*** *Adaptive Synthetic Sampling Approach;* ***SMOTE:*** *Synthetic Minority Oversampling Technique;* ***RUS:*** *Random Under-Sampling;* ***TL:*** *Tomek Links;* ***OSS:*** *One sided selection;* ***ENN:*** *Edited Nearest Neighbor;* ***RF:*** *Random Forest;* ***XGBM:*** *Extreme Gradient Boosting Machine;* ***MMCC:*** *Multi-class Matthews Correlation Coefficient. Bold values present the highest comparison metric in a column*

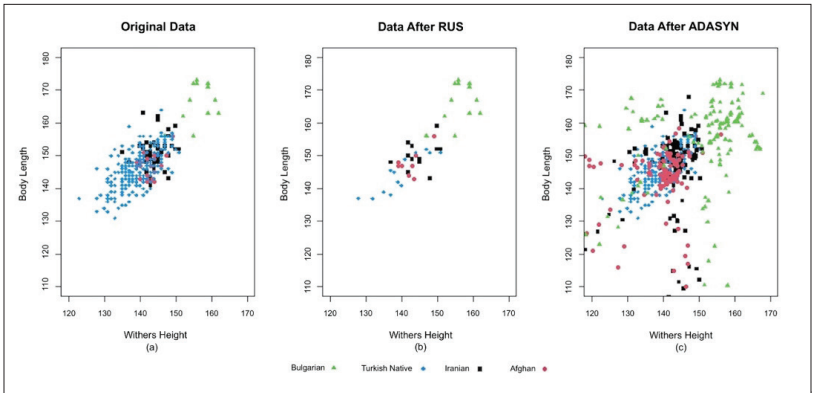

Fig 2. Distributions of the classes belong to the original data, data balanced with random undersampling, and data balanced with adaptive synthetic sampling approach

RUS: Random Under-Sampling; ADASYN: Adaptive Synthetic Sampling Approach

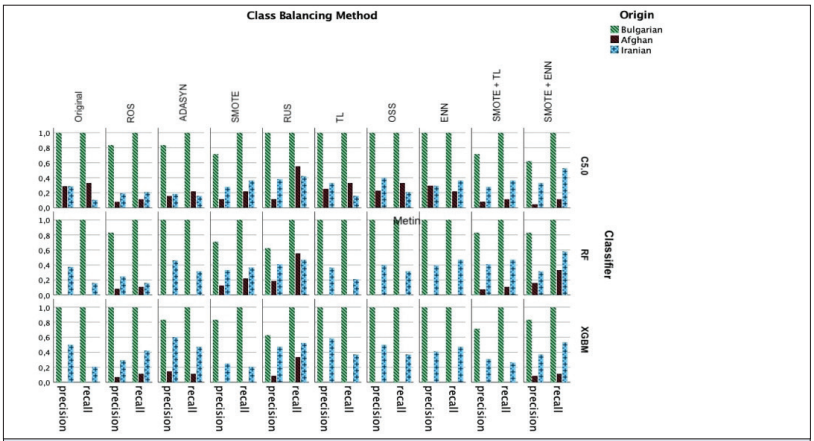

Fig 3. Precision and recall values of minority classes for each balancing method and classifier. *(ROS: Random Over-Sampling; ADASYN: Adaptive Synthetic Sampling Approach; SMOTE: Synthetic Minority Oversampling Technique; RUS: Random Under-Sampling; TL: Tomek Links; OSS: One sided selection; ENN: Edited Nearest Neighbor; RF: Random Forest; XGBM: Extreme Gradient Boosting Machine)*

distribution of the classes can be obtained. The highest number of observations were seen in *Fig. 2-c*, where the distribution of the classes seemed to be more complicated.

Fig. 3 displays the precision and recall results of the minority classes for each class balancing method and classifier. All classifiers that were trained on different balanced data sets had superior prediction performance on Bulgarian horses. It was followed by Iranian and Afghan horses, respectively. Classifiers presented quite poor prediction performance on Afghan horses. Precision and recall values of Afghan horses were found to be quite related to the non- computable results in *Table 3*.

Discussion

In this study we compared some well-known class balancing methods and tried to improve performances of classifiers on classification of pacing horses according to their origins in Türkiye. Class balancing methods, that were categorized under three categories, applied on the training data sets before training the classifiers. Classifiers provided the most successful prediction performance on the data set balanced with RUS according to G-mean and MMCC metrics, where F1 score indicated that best predictive performance was on the data set balanced with ADASYN. Using RUS and ADASYN increased the predictive performance of the classifiers with respect to original data set. Surprisingly, RUS, which was usually regarded as an inefficient method, produced comparable results to more complex methods. This result was consistent with some previous studies. In their study, Drummond and Holte [37] draw some conclusions about how balancing methods enhance the performance of the C4.5 algorithm. They compared over and under-sampling methods using cost curves and found that under-sampling methods produce better results. Ling and Li [38] also compared over and under-sampling for boosted C4.5 and conclude that under-sampling provides superior results, though over-sampling produces nearly as well. On the other hand, there are some studies in the literature that favor oversampling methods. On several data sets, Batista et al.[12] compared the effects of over-sampling, under-sampling, and combined methods. They concluded that over-sampling methods generally outperform the competition. They reported that over-sampling methods outperformed than the other ones in general. Even they proposed some combined methods such as SMOTE+ENN and SMOTE+TL to the literature, concluding that random-oversampling methods provided more meaningful results. Japkowicz and Stephen [14] made a systematic study on class imbalance study with using artificial data sets. They compare various over-sampling and under-sampling methods and conclude that over-sampling had a better way to reduce error rate.

In this study, we also compared some tree-based classifiers. Although RF provided the best prediction performance with respect to G-mean or XGBM provided the highest F1 score, it should be noted that C5.0 algorithm did not produce any NaN results *(Table 3)*. Therefore, we can call C5.0 as the best classifier in the study since it worked well in every scenario. The performance of the classifiers had been negatively impacted by the fact that Afghan horses were not only a minority class but also had a data set that overlaps with Iranian and Turkish Native horses *(Fig. 2-a)*. In the study, although the Bulgarian horses were relatively few and contributed little to the formation of the classifiers, the predictability rate was higher in all scenarios. This supports the idea that in a classification problem with an imbalanced data, having clear decision boundaries can overcome the issue of working with different sample sizes [9,12].

There is one undersampling and one oversampling method that is observed to be the most successful in improving

the prediction performance in the classification problem performed with the data set used in this study, since the highest comparison metrics were observed on the data sets balanced with these methods. While the highest G-mean and MMCC values were observed with the RUS-balanced data set, the highest F1 score value was observed on the data set balanced with ADASYN. On the contrary, the lowest G-mean and MMCC values were seen in the data set balanced with ROS, while the lowest F1 score was seen on the data set balanced with RUS. While G-mean and MMCC were in agreement, F1 score provided results in the opposite direction. In some studies, containing binary or multiclass imbalanced data problems using MCC and MMCC over F1 score were highly recommended. They concluded that F1 score might lead some biased results on imbalanced data scenarios [32,39].

Following the selection of the best-balanced data set and classifier combination, the morphological traits that had contributed the most to the training of the classifiers were evaluated. Body length, withers height, chest circumference, and rump height were determined as predictors that had major roles in classifying pacing horses *(Fig. 1)*. They were followed by chest depth, cannon bone circumference and head length. When the distribution of the classes drawn with body length and withers height were examined *(Fig. 2-a)*, it was observed that Turkish Native class had the lowest measurements, while the Bulgarian horses had the largest. Iranian and Afghan horses, that present overlapping, obviously had similar morphological traits which was supported by Yüceer et al.[16]'s study.

This study is important as it applies machine learning techniques, which are rising in popularity, to the practice of veterinary medicine. This work is useful for several reasons, including the comparison of methods that can be used to solve the imbalance problem across classes when using classification-based machine learning algorithms, as well as the possibility that it will serve as a basis for further research. Also, the Turkish Ministry of Agriculture and Forestry has prioritized breed registration studies in recent years. There are numerous registered breeds among animals other than horses. For the registration of horse breeds, a scientific data and a pedigree are required. In this study, morphological traits were used to classify the pacing horses in Türkiye, including native, Iranian, Afghan, and Bulgarian horses. As a result, the phenotypic diversity of the pacing horses raised in various Turkish regions was identified, and it was found that body length, withers height, chest circumference, and rump height were seemed to be the primary used predictors during training the classifiers. Consequently, this study generated crucial data that can be applied to research on breed registration. Additionally, it might be used for a second-order validation task.

In conclusion, loss of prediction accuracy in a multi-class problem is related to both the presence of numerous minority classes and situations of class overlapping. Class balancing methods can be used to overcome these issues, which were also present in our data set. In comparison to the original imbalanced data set, superior prediction performance results were obtained on the data sets balanced with RUS and ADASYN. Future research may investigate various methods that address the issue of imbalanced data to enhance the classification of pacing horses.

Availability of Data and Materials

The dataset used in the study is available from the corresponding author (H. Özen) on reasonable request.

Funding Support

This study was not financially supported.

Competing Interests

The authors have no conflicts of interest to declare.

Ethical Approval

There is no need for ethical committee approval in the dataset used for this study given that no interventional procedure was involved and only data on morphological traits were used.

Author Contributions

H.Ö. and D.Ö. designed the study and made literature research. H.Ö. and D.Ö. did methodological work and wrote the paper. B.Y.Ö. and C.Ö collected the data set and made contributions to results and discussion.

References

1. Cihan P, Gokce E, Kalipsiz O: A review of machine learning applications in veterinary field. *Kafkas Univ Vet Fak Derg,* 23 (4): 673-680, 2017. DOI: 10.9775/kvfd.2016.17281

2. Burti S, Zotti A, Bonsembiante F, Contiero B, Banzato T: A machine learning-based approach for classification of focal splenic lesions based on their CT features. *Front Vet Sci,* 9:872618, 2022. DOI: 10.3389/fvets.2022.872618

3. Gouda HF, Hassan FA, El-Araby EE, Moawed SA: Comparison of machine learning models for bluetongue risk prediction: A seroprevalence study on small ruminants. *BMC Vet Res,* 18:394, 2022. DOI: 10.1186/s12917-022-03486-z

4. Reagan KL, Deng S, Sheng J, Sebastian J, Wang Z, Huebner SN, Wenke LA, Michalak SR, Strohmer T, Sykes JE: Use of machine-learning algorithms to aid in the early detection of leptospirosis in dogs. *J Vet Diagn Invest,* 34 (4): 612-621, 2022. DOI: 10.1177/10406387221096781

5. Cihan P, Gokce E, Kalipsiz O: A review on determination of computer aid diagnosis and/or risk factors using data mining methods in veterinary field. *Atatürk Üniv Vet Bil Derg,* 14 (2): 209-220, 2019.

6. Nagavelli U, Samanta D, Chakraborty P: Machine learning technology-based heart disease detection models. *J Healthc Eng,* 2022:7351061, 2022. DOI: 10.1155/2022/7351061

7. Xiong T, Ma Z, Li Z, Dai J: The analysis of influence mechanism for internet financial fraud identification and user behavior based on machine learning approaches. *Int J Syst Assur Eng Manag,* 13 (3): 996-1007, 2022. DOI: 10.1007/s13198-021-01181-0

8. Kaur H, Pannu HS, Malhi AK: A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Comput*

Surv, 52 (4): 1-36, 2019. DOI: 10.1145/3343440

9. Vuttipittayamongkol P, Elyan E, Petrovski A: On the class overlap problem in imbalanced data classification. *Knowl Based Syst,* 212:106631, 2021. DOI: 10.1016/j.knosys.2020.106631

10. Krawczyk B: Learning from imbalanced data: Open challenges and future directions. *Prog Artif Intell,* 5 (4): 221-232, 2016. DOI: 10.1007/s13748-016-0094-0

11. Fernández A, López V, Galar M, Del Jesus MJ, Herrera F: Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowl Based Syst,* 42, 97-110, 2013. DOI: 10.1016/j.knosys.2013.01.018

12. Batista GE, Prati RC, Monard MC: A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor,* 6 (1): 20-29, 2004. DOI: 10.1145/1007730.1007735

13. Sáez JA, Krawczyk B, Woźniak M: Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognit,* 57, 164-178, 2016. DOI: 10.1016/j.patcog.2016.03.012

14. Japkowicz N, Stephen S: The class imbalance problem: A systematic study. *Intell Data Anal,* 6 (5): 429-449, 2002. DOI: 10.3233/IDA-2002-6504

15. Caglayan T, Inal S, Garip M, Coskun B, Inal F, Gunlu A, Gulec E: The determination of situation and breed characteristics of Turkish Rahvan horse in Turkey. *J Anim Vet Adv,* 9 (4): 674-680, 2010. DOI: 10.3923/javaa.2010.674.680

16. Yüceer B, Özarslan B, Özbeyaz C: Phenotypic diversity between pacing horses in Turkey. *Ankara Univ Vet Fak Derg,* 63 (2): 195-199, 2016. DOI: 10.1501/Vetfak_0000002729

17. Sha L, Raković M, Das A, Gašević D, Chen G: Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. *IEEE Trans Learn Technol,* 15 (4): 481-492, 2022. DOI: 10.1109/TLT.2022.3196278

18. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res,* 16, 321-357, 2002. DOI: 10.1613/jair.953

19. He H, Bai Y, Garcia EA, Li S: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. **In,** *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, June 1-6, Hong Kong, 2008.

20. Tomek I: Two modifications of CNN. *IEEE Trans Syst Man Cybern*, 6 (11): 769-772, 1976. DOI: 10.1109/TSMC.1976.4309452

21. Kubat M, Matwin S: Addressing the curse of imbalanced data sets: One-sided sampling. **In**, *Proceedings of the 14th International Conference on Machine Learning*. 8-12 July, San Francisco, United States, 1997.

22. Wilson DL: Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans Syst Man Cybern Syst,* 2 (3): 408-421, 1972. DOI: 10.1109/TSMC.1972.4309137

23. Quinlan JR: C4. 5: Programs for Machine Learning. 81-91, Morgan Kaufmann Publishers, San Mateo, 1993.

24. Breiman L: Random forests. *Mach Learn,* 45, 5-32, 2001. DOI: 10.1023/A:1010933404324

25. Schonlau M, Zou RY: The random forest algorithm for statistical learning. *Stata J,* 20 (1): 3-29, 2020. DOI: 10.1177/1536867X20909688

26. Chen T, Guestrin C: Xgboost: A scalable tree boosting system. **In,** *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 13-17 August, San Francisco, 2016.

27. Kuhn M, Johnson M: Classification Trees and Rule-Based Models. **In:** Applied Predictive Modeling. 369-413, Springer, New York, 2014.

28. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, Team RC: Package 'caret'. *R J,* 223 (7):2020.

29. Team R: A language and environment for statistical computing. (Version 4.2.2)[Computer software], Vienna, Austria, 2022.

30. Team R: RStudio: Integrated development environment for R. (Version: 2022.07.2+576)[Computer software], *Boston, MA,* 2022.

31. Branco P, Ribeiro RP, Torgo L: UBL: An R package for utility-based learning. *arXiv preprint arXiv:160408079*, 2016.

32. Tanha J, Abdi Y, Samadi N, Razzaghi N, Asadpour M: Boosting methods for multi-class imbalanced data classification: An experimental review. *J Big Data,* 7 (1): 1-47, 2020. DOI: 10.1186/s40537-020-00349-y

33. Cihan P, Kalipsiz O, Gokce E: Yenidoğan kuzularda bilgisayar destekli tanı. *Pamukkale Univ Muh Bilim Derg*, 26(2): 385-391, 2020. DOI: 10.5505/pajes.2019.51447

34. Japkowicz N: Assessment metrics for imbalanced learning. **In** Imbalanced Learning: Foundations, Algorithms, and Applications. 187-206, Wiley IEEE Press, 2013.

35. Sun Y, Kamel MS, Wang Y: Boosting for learning multiple classes with imbalanced class distribution. **In,** *Proceedings of the 6th International Conference on Data Mining*. 15-18 December, Hong Kong, 592-602, 2006.

36. Halimu C, Kasem A, Newaz SS: Empirical comparison of area under ROC curve (AUC) and Mathew correlation coefficient (MCC) for evaluating machine learning algorithms on imbalanced datasets for binary classification. **In,** *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing*. 25-28 January, Da Lat, Vietnam, 2019.

37. Drummond C, Holte R: C4. 5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. **In,** *Proceedings of the ICML Workshop on Learning from Imbalanced Datasets*. 21 August, Washington DC, 2003.

38. Ling CX, Li C: Data mining for direct marketing: Problems and solutions. **In,** *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*. 27-31 August, New York, 73-79, 1998.

39. Chicco D, Jurman G: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom*, 21:6, 2020. DOI: 10.1186/s12864-019-6413-7