IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 6, 2019, accepted May 29, 2019, date of publication June 7, 2019, date of current version June 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2921591

An Evolutionary UnderBagging Approach to Tackle the Survival Prediction of Trauma Patients: A Case Study at the Hospital of Navarre

JOSE ANTONIO SANZ[1,2], **MIKEL GALAR**[1,2], **(Member, IEEE),**
HUMBERTO BUSTINCE[1,2], **(Senior Member, IEEE), AND TOMAS BELZUNEGUI**[3]

[1]Department of Statistics, Computer Science, and Mathematics, Universidad Publica de Navarra, 31006 Navarra, Spain
[2]Institute of Smart Cities, Universidad Publica de Navarra, 31006 Navarra, Spain
[3]Complejo Hospitalario de Navarra, Instituto de Investigación de Navarra, Navarrabiomed, IDISNA (Instituto de Investigación de Navarra), 31008 Navarra, Spain

Corresponding author: Jose Antonio Sanz (joseantonio.sanz@unavarra.es)

This work was supported in part by the Spanish Ministry of Science and Technology under Project TIN2016-77356-P (AEI/FEDER, UE), in part by the Network Project under Grant TIN2014-56381-REDT, and in part by the Health Department of the Government of Navarre under Project PI-019/11.

ABSTRACT Survival prediction systems are used among emergency services at hospitals in order to measure their quality objectively. In order to do so, the estimated mortality rate given by a prediction model is compared with the real rate of the hospital. Hence, the accuracy of the prediction system is a key factor as more reliable estimations can be obtained. Survival prediction systems are aimed at scoring the severity of patients' injuries. Afterward, this score is used to estimate whether the patient will survive or not. Luckily, the number of patients who survive their injuries is greater than that of those who die. However, this degree of imbalance implies a greater difficulty in learning the prediction models. The aim of this paper is to develop a new prediction system for the Hospital of Navarre with the goal of improving the prediction capabilities of the currently used models since it would imply having a more reliable measurement of its quality. In order to do so, we propose a new strategy to conform an ensemble of classifiers using an evolutionary under sampling process in the bagging methodology. The experimental study is carried out over 462 patients who were treated at the Hospital of Navarre. Our new ensemble approach is an appropriate tool to deal with this problem as it is able to outperform the currently used models by the staff of the hospital as well as several state-of-the-art ensemble approaches designed for imbalanced domains.

INDEX TERMS Ensembles, evolutionary algorithms, imbalanced classification, survival prediction, trauma.

I. INTRODUCTION

Severe trauma patients are persons who have several injuries caused by energy interchanges [1] such as car crashes or falls. The goal of the emergency services is to save as many persons as possible and to try to make them have the best possible life quality after their recovery, as well. The later fact is not only beneficial for the patients but it also implies a reduction in the expenses derived from the subsequent treatments prescribed to these patients.

The survival rate of trauma patients is a good quality indicator of the emergency services. However, it is not an objective measure because the severity of the injuries of the treated patients can vary depending on the hospital and/or the period of time. For example, one hospital may have a survival rate of 97%, whereas another one could stand with 85%. Nevertheless, such a great difference at first glance may led to draw incorrect conclusions, since the former hospital may receive less seriously injured patients. Consequently, it is important to develop tools allowing one to objectively measure the quality of emergency services by making use of the severity of the injuries of the patients to properly assess its survival status. For this reason, survival prediction systems were developed in such a way that a number representing the survival probability of a patient was given, which was thereafter converted to a class of survival or death. In this manner, it is possible to compute standardized mortality rate [2] dividing the real mortality rate by the predicted mortality

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

rate, which allows one to determine the quality of emergency service of the hospitals. If the real rate is close to the predicted one (standardized mortality rate close to 1), the emergency service would be working properly. In summary, a survival prediction system is used as a normalization factor in order to be able to compare different hospitals or the same hospital over time. Consequently, the more accurate the prediction is, the better the measurement of the quality of the service will be.

Nowadays there are standard methods that are used to predict the survival status of trauma patients. One of the most applied ones is the TRauma - Injury Severity Score (TRISS) [3], which was developed in USA with the data obtained in the major trauma outcome study. This method is based on a logistic regression model as most of the current prediction systems do [4], [5]. However, the usage of a general system for every hospital may be biased when applied to different hospitals from different regions. This is why developing new models using the data for each hospital can improve the results given by TRISS, obtaining a more reliable measurement tool for the corresponding hospital.

The survival prediction of trauma patients is a classification problem [6], since there are only two possible outcomes in the system: *survive* and *die*. Nowadays the application of soft computing techniques is widely accepted to tackle different types of classification problems [7]–[10]. Fortunately, the number of severe trauma patients who survive to their injuries is greater, to a large degree, than that of those patients who die. In data mining, this problem is known as the class imbalanced problem [11], [12], since there are more instances belonging to one class (majority class, survive) than to the other (minority class, die). This problem is a challenge for machine learning techniques [13] because classifiers usually tend to predict the majority class for all the instances, and consequently most of the time they fail the prediction of the instances belonging to the minority class.

Techniques applied to face imbalanced classification problems can be categorized in three main groups: 1) internal approaches [14], [15], which create algorithms or modify existing ones, 2) external techniques [11], [16], which add a preprocessing step where the data is sampled (balanced) before the learning process, and 3) cost-sensitive methods [17], which consider the two former to take into account the misclassification costs in the learning process. In the last years the usage of ensembles of classifiers is emerging to tackle this problem [18]–[20]. Ensembles are aimed at increasing the performance of single classifiers by learning several of them and, when classifying new instances, querying all of them and combining their outputs to determine the class. However, for the sake of dealing with imbalanced problems they have to be specifically designed. On this regard, there is a positive synergy between sampling techniques and ensemble-based algorithms, since they usually enhance the results obtained when applying sampling techniques before learning a single classifier [12].

This work is aimed at developing a new prediction system adapted to the features of the trauma patients treated at the Hospital of Navarre (Spain). Consequently, the measurement of the quality of the emergency service of this hospital can be improved by comparing the real status of these patients with the predictions of the system. To do so, we propose the usage af an ensemble based model, instead of a single prediction system, combined with sampling techniques to deal with the imbalanced data problem. Specifically, we propose a new technique to conform an ensemble of classifiers (C4.5 decision trees [21] are used in this paper) by combining the Evolutionary Under Sampling (EUS) [22] algorithm with the bagging methodology. That is, the generation of each bag is carried out by applying EUS. Consequently, the quality of each bag may be increased so that it leads to learning better base classifiers whereas maintaining their diversity (due to the usage of a specific mechanism), which can enhance the overall performance of the ensemble. Moreover, this is a novel approach as it is the first method combining EUS and bagging. Although there is an ensemble making use of EUS and boosting [23], we will show that our new proposal enhances its results in this problem. All in all, the main novelties of this work are:

1) The definition of a new methodology to build an ensemble of classifiers for imbalanced classification problems. Specifically, it creates a bagging based ensemble where EUS is applied in each bag to select the most important patients of the survive class.
2) The application for the first time of ensembles of classifiers considering sampling techniques in their construction to tackle the survival prediction of trauma patients.

The experimental study has been conducted over the patients stored in the *Major Trauma Registry of Navarre* (MTRN) [24]. Specifically, the MTRN is composed of 462 patients who were treated at the emergency services of the Hospital of Navarre in 2011 and 2012. We compare the results of our proposal with those provided by the following single models:

- TRISS and the Mortality Prediction Model of Navarre (MPMN) [5], since they are applied at the Hospital of Navarre.
- The cost sensitive version of the C4.5 decision tree [25] as it is a well-known technique for imbalanced domains.

Moreover, we also consider in the comparison easy-ensemble [19] as well as with several ensemble methods based both on bagging and boosting designed for imbalanced domains [12]. The quality of the results is measured using three well-known metrics for imbalanced domains like the area under the ROC curve (AUC) [26], the geometric mean (GM) [27], which quantifies the balance between specificity and sensitivity, and the F-measure [28]. The results are supported by a proper statistical study, which is conducted using the Mann-Whitney's U statistical test [29].

The remainder of this work is organized as follows: Section II describes the problem tackled in this work. In Section III the necessary concepts about imbalanced classification problems are introduced including related ensemble approaches. Next, our new proposal is described in detail in Section IV. The obtained results and the corresponding analysis are shown in Section V and finally, the main conclusions are drawn in Section VI.

II. SURVIVAL STATUS PREDICTION OF SEVERE TRAUMA PATIENTS: PROBLEM DESCRIPTION

Trauma patients are persons suffering from several serious injuries, which imply a risk for their life. It is one of the most frequent causes of death for people under 40 and it also implies high economic expenses for health centers [30]. These patients usually follow an established medical treatment, and therefore there is a relation between the therapeutic measures taken and the survival status of the patients, which can take only two values: *survive* or *die*.

Survival prediction systems are applied to convert the severity status of these patients into a probability representing the likelihood of being able to survive, which can be straightforwardly transformed into the two mentioned classes: survive and die. These measurements can be used to compare two health centers objectively, taking into account the severity of the patients they have to deal with. In summary, if a hospital is able to safe the life of more patients than those that were predicted by the model, it would be classified as a good quality hospital because it would be working better than the standard.

The goal of any hospital control system is to perform a continuous and measurable improvement of the treatments applied to patients. With this aim, the information obtained from all the severe trauma patients treated at health centers is stored in a database named Major Trauma Registry (MTR) [24]. A MTR is a precise and complete source of information that allows one to continuously monitor the assistance process in the trauma center units. A well-designed MTR helps hospital managers in analyzing the information trying to discover facets that can be changed, aimed at improving the quality of life of the survivors and coordinating the different services involved in care units. Both the monitor and quality control processes have allow the mortality and disability rates of these patients to be reduced in developed countries in the last years [31].

The emergency department of the Hospital of Navarre (Spain) conducted a study that allowed them to develop and validate the MTR of Navarre (MTRN) [24]. This registry is based on the Utstein model [32], which determines the features to be collected (a total of 53). Some of them are easily obtained like the age or the gender of the patients, whereas the other ones are based on the severity of the injuries of the patients such as the *Injury Severity Score* (ISS) [33], the *New Injury Severity Score* (NISS) [34] or the *Revised Trauma Score* (RTS) [35].

We have to point out that not all the severe trauma patients are stored in the MTRN. There exist the following five exclusion criteria:

TABLE 1. Profile of the patients stored in the major trauma registry of Navarre.

Variable	Number of patients	Die	Survive
Total patients	462	94 (20.3%)	368 (79.7%)
Age	53.1 (22.9)	66.4 (22.1)	49.7 (21.7)
Gender			
Male	324 (70.1%)	60 (18.5%)	264 (81.5%)
Female	138 (29.9%)	34 (24.6%)	104 (75.4%)
Premorbid conditions			
Healthy patient	294 (63.6%)	36 (12.2%)	258 (87.8%)
Mild systemic disease	129 (28%)	41 (31.8%)	88 (68.2%)
Severe systemic disease	39 (8.4%)	17 (43.6%)	22 (56.4%)
Type of injury			
Blunt	442 (95.7%)	93 (21%)	349 (79%)
Penetrating	20 (4.3%)	1 (5%)	19 (95%)
Mechanism of injury			
Traffic	192 (41.6%)	34 (17.7%)	158 (82.3%)
Shot by handgun or stabbed by knife	14 (3.0%)	1 (7.1%)	13 (92.9%)
Low energy fall	153 (33.1%)	42 (27.4%)	111 (72.6%)
High energy fall	66 (14.3%)	15 (22.7%)	51 (77.3%)
Other	37 (8%)	2 (5.4%)	35 (94.6%)
Intention of Injury			
Accident (unintentional)	422 (91.4%)	84 (19.9%)	338 (80.1%)
Self-inflicted	21 (4.5%)	6 (28.6%)	15 (71.4%)
Assault	19 (4.1%)	4 (21%)	15 (79%)
Physiological scores			
RTS upon arrival of EMS personnel at scene	7.3 (1.1)	6.3 (1.6)	7.5 (0.8)
T-RTS upon arrival of EMS personnel at scene	11.3 (1.3)	10.2 (1.9)	11.6 (0.9)
RTS upon arrival in A&E/hospital	6.9 (1.5)	5.5 (1.8)	7.3 (1.1)
T-RTS upon arrival in A&E/hospital	11.0 (1.7)	9.4 (2.1)	11.4 (1.2)
Anatomically based severity scores			
ISS	20.5 (8.9)	27.9 (9.9)	18.6 (7.6)
NISS	27.3 (10.3)	37.3 (12.2)	24.7 (7.9)
Analytical parameter			
Coagulation: INR	1.2 (0.7)	1.4 (1.1)	1.1 (0.5)
Prehospital intubation			
No	414 (89.6%)	70 (16.9%)	344 (83.1%)
Yes	48 (10.4%)	24 (50%)	24 (50%)

1) The value of the NISS feature is less than 15.
2) The period of time among the injury and the hospital admission is greater than 24 hours.
3) The patient was drowned.
4) The patient was hanged.
5) The patient was burnt.

Specifically, the MTRN stores data of 462 patients collected between 2011 and 2012,[1] 368 of them survived to their injuries whereas the remainder 94 died. Consequently, it is an imbalanced classification problem as there is a larger number of patients who survive than that of those who die. In Table 1 we show a summary of the profile of the patients stored in the MTRN.

A. RELATED WORKS

The comparison of the results achieved by different health institutions at any level (regional, national or international) allows one to enhance the data collection and the patient survival [4], [36]. Soft computing techniques are usually considered to do so. The best example is the standard method in this domain, that is, the Trauma and Injury Severity Score (TRISS) [3]. This method is based on a logistic regression and its input variables are the ISS [33], the RTS [35] and the age, which is binarized. However, the performance of TRISS can be enhanced by learning the model parameters according to the features of new patients as it is currently done in this field [37]–[39]. New adjustments for the TRISS model have been recently published like [40]. There are also recent papers where we can find comparisons among the TRISS methodology and new prediction systems like TARN

[1] At the moment when the staff of the Hospital of Navarre provided us the data.

and NORMIT [41] or models focused on specific segments of the population like geriatric trauma patients [42]. The second version of the Revised Injury Severity Classification (RISC II) [4] was developed in order to deal with the limitations detected in its first version (RISC) [43]. This model considers laboratory values like base deficit, haemoglobin's concentration and thromboplastin time for the first time, as well as medical interventions such as cardiopulmonary resuscitation (CPR) [44].

The staff of the Hospital of Navarre made a review of prediction techniques [45] and developed their own model [5], which is named as Mortality Prediction Model of Navarre (MPMN). This model is also a logistic regression whose input variables are the age, the RTS, the NSS and the previous morbidity. The performance of MPRN is similar to that of RISC II for the patients stored in the MTRN [46]. Furthermore, trying to provide a more interpretable model they made usage of decision trees as well as sampling techniques to tackle the imbalanced data in [47]. Finally, they also proposed to apply a multiple classifier system in [48] for improving the performance of individual models.

As it can be observed, all the methods but the last one relies on the usage of a single classifier to tackle the survival prediction problem and consequently, whether ensembles of classifiers would improve the performance of the system or not remains as an open question, which we aim to answer in this paper.

III. IMBALANCED CLASSIFICATION PROBLEMS

This section is aimed at introducing the background about the class imbalance problem besides the proper performance metrics for this problem (Section III-A) and describing the ensemble methods related to our proposal (Section III-B).

A. CLASS IMBALANCE PROBLEM AND PERFORMANCE METRICS

Before defining the imbalance classification problems, we recall the concept of supervised classification. A classification problem consists in learning a function called *classifier* that is able to predict the class of new incoming examples. To do so, a training set $\mathcal{D}_T$ composed of P labeled examples $x_p = (x_{p1}, \ldots, x_{pn}), p = \{1, \ldots, P\}$, where x_{pi} is the value of the i-th variable ($i = \{1, 2, \ldots, n\}$) of the p-th training example. Each example belongs to a unique class $y_p \in \mathbb{C} = \{C_1, C_2, \ldots, C_m\}$, where m is the number of classes of the problem.

An imbalanced classification problem [11], [12] is a classification problem where the number of examples belonging to the different classes is considerably different. That is, the class distribution is not uniform. When tackling two class problems, the class having the largest number of examples is known as majority class (or negative) and the other class is known as minority class (or positive). The Imbalanced Ratio (IR) [49], is computed by dividing the number of examples belonging to the majority class by that of the minority class.

TABLE 2. Confusion matrix for a two class problem.

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

A key point when tackling classification problems is the measurement of the system's performance. Classically, the usage of the percentage of correctly classified examples (accuracy rate) is used to asses the quality of the classifiers. However, in imbalanced domains it is no longer a proper measure, since the majority class clearly dominates this metric. Therefore, to measure the quality of the classifiers in imbalanced classification problems, the accuracy on each class has to be taken into account simultaneously. There are several proper metrics, which are constructed from the confusion matrix (Table 2), which stores the number of correctly and incorrectly classified examples in each class.

From this matrix, different measures can be computed to perform the evaluation in an imbalanced framework:

- *True positive rate:* It is also known as *recall* and it is the percentage of positive instances correctly classified, which is computed as $TP_{rate} = \frac{TP}{TP+FN}$
- *True negative rate:* It is the percentage of negative instances correctly classified, which is computed as $TN_{rate} = \frac{TN}{FP+TN}$
- *False positive rate:* It is the percentage of negative instances misclassified, which is computed as $FP_{rate} = \frac{FP}{FP+TN}$
- *False negative rate:* It is the percentage of positive instances misclassified, which is computed as $FN_{rate} = \frac{FN}{TP+FN}$

However, these measures on their own are still inadequate because they do not consider both classes at the same time. In this work, we have selected three performance metrics that are suitable for this domain.

The first one is the geometric mean (GM) [27], which computes the geometric mean between the accuracy obtained in each class as it is shown in (1).

$$GM = \sqrt{TP_{rate} \cdot TN_{rate}} \tag{1}$$

The second performance metric is the Area Under the ROC Curve (AUC) [26]. It allows one to take into account the balance between TP_{rate} and FP_{rate}, which tries to show that increasing the number of true positives without also increasing the number of false positives is not possible for any classifier. This performance metric is obtained applying (2).

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{2}$$

The third one is the F-measure [28] that is defined as the harmonic mean between precision, which is the percentage of correctly classified instances of those predicted as positive

($\frac{TP}{TP+FP}$), and recall (TP_{rate}) as shown in (3).

$$F - measure = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3}$$

B. ENSEMBLE METHODS TO DEAL WITH IMBALANCED CLASSIFICATION PROBLEMS

The problem of imbalanced classification is present in many real-world problems. In the last years, there is an increasing number of solutions using ensembles of classifiers to deal with it. An ensemble is a classifier that, in turn, is composed of several classifiers [12]. Their objective is to enhance the performance of single classifiers by learning a group of them, which are known as base classifiers. To classify new instances, all the base classifiers are queried and their outputs are aggregated to determine the class.

In [12] authors review the state-of-the-art in ensemble-based solutions for imbalanced problems. Specifically, a taxonomy was proposed and the quality of ensemble approaches was validated by an extensive experimental study. In this section we briefly describe the methods we have used in the comparative study.

- Boosting-based ensembles: AdaBoost [50], which is the most representative approach, uses the entire dataset to train serially a set of classifiers. It assigns weights to the instances so that in each iteration, the learning process of the base classifier is focused on the most difficult instances. The update of the weight increases the weight of those instances misclassified by the base classifier generated whereas it decreases that of the correctly classified ones. Additionally, AdaBoost also assign weights to the base classifiers according to their performance. The classical AdaBoost algorithm has been modified to tackle imbalanced problems by carrying out a sampling process in each iteration so that only the selected instances are used in the learning process. We have considered the following three methods:
  1) RUSBoost [18]: It applies a random undersampling process in order to remove instances belonging to the majority class. Then, the weights of the remainder instances are normalized with respect to their sum.
  2) EUSBoost [23]: This technique is aimed at improving the behavior of RUSBoost by avoiding the randomness. To do it, EUS is applied to select the instances of the majority class considering both the performance and the diversity in the fitness function. Finally, the distribution of the weights is also updated according to the selected instances.
  3) SMOTEBoost [51]: This method enlarges the dataset by including new instances of the minority class applying the SMOTE algorithm [52], which creates synthetic instances using an interpolation procedure. The weights of the new instances are proportional to the total number of instances in the enlarged dataset and they will be the same in all the iterations. On the other hand, the weights of the original instances are normalized so that they form a distribution with the new instances.
- Bagging-based ensembles: approaches in this group are based on the concept of bootstrap aggregating [53]. That is, different classifiers are trained using bootstrapped replicas of the original training dataset, which usually have the same size than the original dataset. The replicas are obtained by randomly drawing (with replacement) instances from the original dataset. The combination of bagging and sampling techniques is usually simpler than with boosting, since it does not require to recompute any kind of weights. When using this hybridization, the bag used to train each base classifier is obtained using the sampling method instead of performing a random selection of the instances.
  1) OverBagging [54]: This method applies a random over sampling process to obtain each bag. As a result, each bag will include all the original instances as well as the replicas of the randomly selected instances of the minority class. In order to boost diversity, instances of the majority class can be resampled.
  2) SMOTEBagging [54]: This technique follows the same schema than OverBagging but instead of applying a random oversampling process it generates new synthetic examples belonging to the minority class applying the SMOTE algorithm [52]. Furthermore, for the minority class, it combines the random resampling process and SMOTE by using a percentage, which is increased over the iterations, determining the amount of instances included by each option. The majority class instances are also randomly resampled to increase the diversity.
  3) UnderBagging [55]: This approach is similar to OverBagging but it uses a random under sampling process instead of an over sampling one. Therefore, the size of each bag is less than that of the ones obtained with OverBagging. It also includes the option of randomly resampling the instances of the minority class.
  4) UnderOverBagging [54]: This method uses undersampling and oversampling. Furthermore, as SMOTEBagging, it also uses a resampling rate but it determines the number of instances taken from each class. Consequently, the first classifiers are trained with a lower number of instances than the last ones, which may imply boosting the diversity.
- Hybrid ensembles: the methods belonging to this group combine both bagging and boosting. The selected approach is EasyEnsemble [19] that applies a double ensemble learning process considering Bagging as the main one, where each bag is balanced including all the instances of the minority class and randomly undersampling the majority one (UnderBagging). Then, for

each bag, the AdaBoost algorithm is applied and consequently, the final model is an ensemble of ensembles.

IV. EUNDERBAGGING: A NEW METHOD TO CONFORM ENSEMBLES BASED ON EVOLUTIONARY UNDER SAMPLING AND BAGGING

In this section we describe our new approach for designing an ensemble of classifiers to tackle imbalanced classification problems like the one faced in this paper. Specifically, our proposal is named EUnderBagging as it combines EUS [22] with a bagging-based ensemble. As we have explained in Section III-B, the classical bagging method randomly selects the instances (with replacement) to create the bag of instances used to learn the base classifier. Our proposal is based on replacing that random selection process (creation of the bootstrapped replica) by the EUS algorithm, which will create in each iteration the bag of instances used to train the corresponding base classifier. Consequently, in first place we describe the EUS method [22] (Section IV-A) and then, the complete EUnderBagging algorithm is described in detail (Section IV-B).

A. EVOLUTIONARY UNDER SAMPLING ALGORITHM

The Evolutionary Under Sampling (EUS) algorithm [22] comes from the application of evolutionary prototype selection in imbalanced classification due to the fact that some of their original features, like the fitness function, can be specifically designed for that problem.

The aim of prototype selection is to select a subset of the instances in the training set in such a way that the nearest neighbor algorithm (1NN) [56] enhances its accuracy rate and lightens its storage requirements. In imbalanced problems obtaining a balanced class distribution becomes more important, since both classes would have the same importance in the learning process. For this reason, the fitness function used by EUS takes into account the class distribution (as explained above, see (5)). The evolutionary process starts selecting at random several subsets of instances that are evolved until one of them cannot be improved in terms of the fitness function, which is the returned solution.

The representation of the solution is a key factor in all the evolutionary algorithms. In EUS the solutions are represented by a chromosome composed of as many genes as instances, where each gene is binary coded to represent the presence (1) or absence (0) of the corresponding instance. To diminish the search space, the number of genes is equal to the number of instances belonging to the majority class. Therefore, the selection process is carried out only over the negative instances and all the instances of the minority class will be always included in the returned subset. The representation of the chromosomes is shown in 4.

$$Chr_{EUS} = (gen_{x_1}, gen_{x_2}, gen_{x_3}, gen_{x_4}, \ldots, gen_{x_{n^-}}), \quad (4)$$

where gen_{x_i} takes the values 0 or 1, indicating whether instance x_i is included or not in the subset, and n^- is the number of majority class instances.

To evaluate the quality of the chromosomes, a fitness function considering both the trade-off between the percentage of instances of both classes and the expected performance when using the selected subset is applied. The resulting fitness function is shown in (5).

$$fitness_{EUS} = \begin{cases} GM - \left|1 - \frac{n^+}{N^-} \cdot P\right| & \text{if } N^- > 0 \\ GM - P & \text{if } N^- = 0, \end{cases} \quad (5)$$

where n^+ is the number of examples of the minority class and N^- is the number of selected examples of the majority class. Consequently, for the subset of instances represented by the chromosome, the division $\frac{n^+}{N^-}$ quantifies the balance of the instances belonging to both classes. GM is the performance, measured in terms of the geometric mean, obtained by the 1NN algorithm considering the *leave-one-out* technique. P is a weight that determines the importance given to the balance part of the equation, whose recommended value is 0.2.

We must point out that the evolutionary algorithm used is the CHC [57]. The crossover operator used in the CHC algorithm for binary coding is the heterogeneous uniform cross-over (HUX), which interchanges half of the different genes in the chromosomes being combined. This crossover is modified in EUS for the sake of obtaining a good reduction rate. Specifically, the probability of including instances is reduced. To do so, when a gene is switched on it can be switched off with a probability, whose recommended value is 0.25.

B. EVOLUTIONARY UNDER BAGGING

The combination between EUS and Bagging is simple, since it consists of applying the EUS algorithm to conform the subset of instances used to learn each base classifier. The pseudo-code of our new approach is shown in Algorithm 1, where it can be seen that for each iteration the corresponding bag is obtained applying EUS, which returns a subset of instances including all the instances from the minority class and those that are selected in the evolutionary process.

Algorithm 1 EUnderBagging

```
Require: Training set $S$: Training set; $T$: Number of iterations; $I$: Weak learner
Ensure: Bagged classifier: $H(x) = sign \sum_{t=1}^{T} h_t(x)$, where
    $h_t \in [-1, 1]$ are the induced classifiers
for $t = 1$ to $T$ do
    $S'$ = EvolutionaryUndersampling(S);
    $h_t \leftarrow I(S')$
end for
```

The idea of EUnderBagging is based on the methodologies of UnderBagging [55] and EUSBoost [23]. In the former, a random under sampling process is applied to build each bag. This randomness implies obtaining diverse subsets of instances, which may lead to construct high performing ensembles when accurate base classifiers are used [58].

However, in imbalanced problems, the random selection may imply that useful instances from the majority class can be skipped for the learning of the base classifiers. Consequently, we think that the application of EUS presents a good trade-off between the two following properties:

- Diversity: the initial population used by the evolutionary algorithm, which is initialized at random, allows the diversity to be partially maintained.
- Selection of important instances: the evolutionary process increments the probability of selecting the important instances of the majority class, which may imply learning a better model in each iteration.

However, although there are random mechanisms in EUS the stochastic nature of the evolutionary process does not provide subsets as diverse as a purely random technique. Consequently, in order to boost the diversity as much as possible, we modify the fitness function as it is done in EUSBoost [23]. The modification is aimed to favor those chromosomes having the best combination of performance, which is measured applying (5), and diversity. To compute the diversity we compare a chromosome and all the subsets of instances used in the previous iterations of the bagging process, since we assume that different subsets of instances will produce diverse models. Specifically, we take the maximum value of the Q-statistic [59] over the chromosome and all the previously used subsets of instances. Consequently, in each iteration, we will take the most different subset of instances with respect to the ones used in the previous iterations. The Q-statistic between two binary vectors (chromosomes) (C_i, C_j), is computed as follows:

$$Q_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}, \tag{6}$$

where N^{ab} represents the number of genes (instances) with value a in the first chromosome and b in the second one. Recall that $a, b \in \{0, 1\}$, since we use a binary representation as mentioned in Section IV-A. When $a = b$ both subsets are including (or not) the instance. The obtained Q-value ranges in $[-1, 1]$, where the value 0 means that both chromosomes are statistically independent whereas large values (negative and positive) means obtained less diverse chromosomes.

All in all, the fitness function used in EUnderBagging is:

$$fitness_{EUS_Q} = fitness_{EUS} \cdot \frac{1.0}{\beta} \cdot \frac{10.0}{IR} - Q \cdot \beta, \tag{7}$$

where $fitness_{EUS}$ is the original fitness function used by EUS (see (5)), IR is the imbalance ratio, Q is the maximum Q-statistic and β is a weighting factor that changes over the iterations as follows:

$$\beta = \frac{T - t - 1}{T}. \tag{8}$$

According to this weight, in the first iterations the importance given to the diversity and the performance is similar whereas in the last iterations more significance is given to the performance and less to the diversity.

Finally, we have to point out that in (7) the Q-statistic is subtracted in order to maximize the diversity. Furthermore, in the first iteration ($t = 1$), the original fitness function of EUS is considered as there are no previous bags of instances used and consequently, it is not possible to compute the Q-statistics.

V. EXPERIMENTAL STUDY

In this section we show the results obtained when using the approaches selected in this study. The experimental framework used to conduct the experiments besides the considered algorithms are introduced in Section V-A. The results as well as their corresponding analysis are given in Section V-B.

A. EXPERIMENTAL FRAMEWORK

The dataset is composed of the information collected from 462 patients that were stored in the MTRN during the years 2011 and 2012. 368 out of the 462 patients survived to their injuries (79.65%) whereas the remainder 94 died (20.35%). Consequently, the IR of the problem is 3.91.

To determine the performance of the classifiers, one of the most used methods is the k-cross validation model (k-FCV). In this work, we have applied a 10 × 10-FCV ($k = 10$). To apply a 10-FCV we first have to split the set of examples in 10 folds having the same number of patients and maintaining the original distribution of the classes. Next, 9 of them are joined to learn the classifier and the remainder one is used to test the quality of the system. This process is repeated 10 times using a different testing fold in each case. Consequently, when the process is ended all the patients will have been used as testing instances once. The whole 10-FCV process is repeated 10 times (obtaining the 10 × 10-FCV) using a different seed each time to perform the splitting. The final result shown in this study is the average among the 100 testing folders. The 10 × 10-FCV allows one to provide robust results as the evolutionary process is carried out 100 times using different data in each run.

In each fold, we consider three widely used performance metrics to measure the performance of the classifiers: the Area Under the ROC Curve (AUC) [26], the geometric mean (GM) [27] as well as the F-measure [28] (we also show the TN_{rate} and the TP_{rate}).

To support the quality of the proposals we apply the non-parametric Mann-Whitney's U statistical test [29] to compare the results of two methods. This method, in first place, sorts the results of both methods in ascending order assigning ranks to the results so that the worst and the best ones receive the ranks 1 and maximum (two times the number of results), respectively. In case of draws, the corresponding ranks are equally assigned. Next, the sum of the ranks is computed for each method. Consequently, if a method is regularly better than the other, the sum of its ranks will be clearly greater than that of the other, which is reflected by a low p-value. Otherwise, if both methods provide similar results, the sum of their ranks will be also similar leading to a large p-value. We have considered 0.1 as the

TABLE 3. Testing results obtained for the sampling techniques with and without being combined with ensembles.

Baseline Sampling Method	Classifier	AUC	GM	F-measure	TN_{rate}	TP_{rate}
RUS	RUS	0.7934 ± 0.0734	0.7873 ± 0.0766	0.6063 ± 0.0976	0.7765 ± 0.0848	0.8103 ± 0.1468
	RUSBoost	0.8174 ± 0.0747	0.8135 ± 0.0764	0.6521 ± 0.1019	**0.8265 ± 0.0665**	0.8082 ± 0.1370
	UnderBagging	**0.8402 ± 0.0711**	**0.8372 ± 0.0715**	**0.6724 ± 0.0994**	0.8159 ± 0.0663	**0.8646 ± 0.1206**
EUS	EUS	0.8115 ± 0.0778	0.8073 ± 0.0788	0.6291 ± 0.1019	0.7843 ± 0.0784	0.8388 ± 0.1345
	EUSBoost	0.8293 ± 0.0721	0.8255 ± 0.0731	0.6509 ± 0.0960	0.7972 ± 0.0647	0.8614 ± 0.1283
	EUnderBagging	**0.8498 ± 0.0648**	**0.8472 ± 0.0650**	**0.6838 ± 0.0912**	**0.8196 ± 0.0601**	**0.8800 ± 0.1090**
SMOTE	SMOTE	0.8143 ± 0.0764	0.8098 ± 0.0814	0.6550 ± 0.1085	0.8387 ± 0.0675	0.7900 ± 0.1381
	SMOTEBoost	0.7816 ± 0.0825	0.7659 ± 0.0962	0.6446 ± 0.1235	**0.9022 ± 0.0543**	0.6610 ± 0.1594
	SMOTEBagging	**0.8220 ± 0.0738**	**0.8180 ± 0.0777**	**0.6570 ± 0.1007**	0.8281 ± 0.0630	**0.8160 ± 0.1381**

lowest level of significance of a hypothesis that results in a rejection.

Regarding the configuration of the different approaches, in first place, we have to point out the we have used the C4.5 decision tree [21] as base classifier for the ensembles and as the model for the approaches using a single classifier. In all the cases we have set the confidence level to 0.25, using the Laplace correction and a minimum of 2 examples per leaf.

According to the recommendations given in [12], the number of base classifiers has been set to 40 and 10 for bagging and boosting-based ensembles, respectively. On the other hand, to make a fair comparison we have used 4 bags using 10 classifiers in each one for EasyEnsemble.

Finally, the configuration of the evolutionary algorithm for those approaches using it is as follows: the populations are composed of 50 individuals, 10.000 iterations, the inclusion probability for the HUX crossover operator is 0.25, the parameter P is set to 0.2 and the Euclidean distance is used in the 1NN algorithm.

For the comparative study we have also considered the cost sensitive version of the C4.5 decision tree ($C45\_CS$) [25], since it is usually applied in the imbalanced context. Moreover, we have also considered the two currently used methods by the staff of the Hospital of Navarre to confirm the quality of our proposal, namely, TRISS and MPMN. Finally, we have also considered our previous proposal where we applied a Multiple Classifier System (MCS) to tackle this problem as well as the TRISS method whose parameters has been learned according the patients stored in the MTRN (TRISS_{Nav}).

We have to stress that for all the methods in the comparison (except TRISS, TRISS_{Nav} and MCS) we have used the same input variables as those used by MPMN, that is, the age, the RTS, the NISS and the previous morbidity.

B. ANALYSIS OF THE PERFORMANCE OF EUNDERBAGGING

The experimental analysis to show the quality of our new approach is driven in four stages:

1) First, we check the usefulness of ensembles of classifiers combining sampling techniques by comparing them versus the sampling technique applied with a single C4.5 decision tree (Section V-B.1).
2) Then, we study the performance of the ensemble methods with sampling techniques belonging to the two families, namely, bagging and boosting (Section V-B.2).
3) Next, the best ensemble with sampling is determined by comparing the best ensemble of each family selected in the previous stage (Section V-B.3).
4) Finally, the quality of the best ensemble is contrasted versus the two regression based models used in the Hospital of Navarre as well as with respect to $C4.5\_CS$ (Section V-B.4).

1) COMPARING ENSEMBLE METHODS VERSUS THEIR SINGLE CLASSIFIER COUNTERPART

According to the ensemble methods described in Section III-B, we can observe that they are combined with three sampling techniques, namely, Random Under Sampling (RUS), Evolutionary Under Sampling (EUS) and SMOTE. Therefore, this section is aimed at comparing each sampling technique (with a single decision tree) versus both its bagging-based ensemble and its boosting-based ensemble. The results in testing of these 9 approaches are introduced in Table 3, where we can find a different classifier in each row and a different performance metric in each column (AUC, GM, F-measure as well as TN_{rate} and TP_{rate}). We also show the standard deviation for each metric, ±, to show the robustness of the approaches. The best result for each sampling technique and each metric is highlighted in **bold-face**.[2]

From these results we can stress the fact the usage of ensembles is highly recommendable, since none of the three sampling techniques achieves the best results in any metric. Furthermore, bagging-based ensembles are providing better results (and more robust as the standard deviation is almost always better) than boosting-based ones with the exception of the TN_{rate} for RUS and SMOTE (but at the cost of a decrease in the TP_{rate}). Finally, we have to point out that the combination of SMOTE and ensembles is providing worse results than combining ensembles with RUS and EUS. So, it seems that the generation of new examples of the minority class (die) is not beneficial and the prediction method works better with the real patients than when including synthetically created ones.

To support the previous findings we conduct an statistical study composed of pairwise comparisons between the sampling technique combined with an ensemble and without them for the three performance metrics. These results are reported in Table 4, where each comparison is shown in a row

[2] The structure of all the tables showing the testing performance is the same as the one explained for Table 3

TABLE 4. Results of the Mann-Whitney's U statistical test to the usage of sampling techniques with esembles (R+) and without them (R−).

Comparison	AUC			GM			F-Measure		
	p-value	(R+)	(R-)	p-value	(R+)	(R-)	p-value	(R+)	(R-)
RUSBoost vs. RUS	0.07^*	107.87	93.13	0.04^*	108.78	92.22	$< 0.01^*$	11.91	89.09
UnderBagging vs. RUS	$< 0.01^*$	118.01	82.99	$< 0.01^*$	118.76	82.24	$< 0.01^*$	117.74	82.26
EUSBoost vs. EUS	0.13	106.73	94.27	0.12	106.85	94.15	0.17	106.12	94.88
EUnderBagging vs. EUS	$< 0.01^*$	115.01	85.99	$< 0.01^*$	115.46	85.54	$< 0.01^*$	115.73	85.27
SMOTEBoost vs. SMOTE	$< 0.01^*$	113.47	87.53	$< 0.01^*$	115.01	85.99	0.40	103.96	97.04
SMOTEBagging vs. SMOTE	0.48	103.36	97.64	0.50	103.25	97.75	0.89	101.08	99.92

TABLE 5. Testing results obtained for boosting-based ensembles.

Classifier	AUC	GM	F-measure	TN_{rate}	TP_{rate}
AdaBoost	0.7413 ± 0.0877	0.7135 ± 0.1128	0.5898 ± 0.1425	**0.9103 ± 0.0442**	0.5722 ± 0.1648
RUSBoost	0.8174 ± 0.0747	0.8135 ± 0.0764	**0.6521 ± 0.1019**	0.8265 ± 0.0665	0.8082 ± 0.1370
EUSBoost	**0.8293 ± 0.0721**	**0.8255 ± 0.0731**	0.6509 ± 0.0960	0.7972 ± 0.0647	**0.8614 ± 0.1283**
SMOTEBoost	0.7816 ± 0.0825	0.7659 ± 0.0962	0.6446 ± 0.1235	0.9022 ± 0.0543	0.6610 ± 0.1594

TABLE 6. Testing results obtained for bagging-based ensembles.

Classifier	AUC	GM	F-measure	TN_{rate}	TP_{rate}
OverBagging	0.8043 ± 0.0748	0.7994 ± 0.0792	0.6450 ± 0.1028	0.8449 ± 0.0581	0.7637 ± 0.1372
SMOTEBagging	0.8220 ± 0.0738	0.8180 ± 0.0777	0.6570 ± 0.1007	0.8281 ± 0.0630	0.8160 ± 0.1381
UnderBagging	0.8402 ± 0.0711	0.8372 ± 0.0715	0.6724 ± 0.0994	0.8159 ± 0.0663	0.8646 ± 0.1206
UnderOverBagging	0.7975 ± 0.0762	0.7905 ± 0.0833	0.6461 ± 0.1087	**0.8628 ± 0.0580**	0.7321 ± 0.1403
EUnderBagging	**0.8498 ± 0.0648**	**0.8472 ± 0.0650**	**0.6838 ± 0.0912**	0.8196 ± 0.0601	**0.8800 ± 0.1090**

and results are grouped in groups of three columns (a group for each metric). For each group it is shown the obtained p-value, the average ranks when using ensembles (R+) and without them (R−). We have to point out that when the obtained p-value is less than 0.01 we show it with the notation < 0.01 and, on the other hand, we stress the p-value with * as super-index when there are statistical differences (p-value less than 0.1).

From the statistical results we can observe that the combination of ensembles and RUS is beneficial for both bagging and boosting approaches, since there are statistical differences. Regarding EUS, its combination with bagging-based ensembles is highly recommended whereas with boosting there are not statistical evidences although the performance is enhanced. Finally, regarding SMOTE, it is suitable when used with boosting techniques (there are statistical differences in AUC and GM) but when it is combined with bagging methods their behavior is similar in all the performance metrics.

2) STUDYING THE BEHAVIOR OF BAGGING AND BOOSTING BASED ENSEMBLE METHODS

Once the usefulness of the usage of ensembles is proven for the problem faced in this paper, we conduct an study to determine the best ensemble using sampling techniques. To do so, we present in Tables 5 and 6, the results provided by boosting-based and bagging-based ensembles, respectively.

From these results we can stress the behavior of two approaches (one for each family). On the one hand, EUSBoost achieves the best results among the boosting-based approaches since the results in terms of AUC and GM are better than those of the remainder approaches whereas in terms of F-measure it almost performs equal to RUSBoost. On the other hand, EUnderBagging is clearly the best choice among the bagging-based methods, since it provides the best results in the three performance metrics (having a large increase versus OverBagging, SMOTEBagging and UnderOverBagging). Moreover, they are the most robust methods of their families according to their standard deviations in terms of AUC, GM and F-measure. Additionally, it is worth mentioning that the combination of under sampling techniques with ensembles is providing better results that the combination with over sampling ones. We can think of two possible reasons why these two methods obtain the best results: 1) the usage of under sampling techniques as we have mentioned in the previous section and 2) EUS (both methods apply it) is the best choice as sampling method, since the selection of the patients of the majority class is driven by an evolutionary algorithm that selects the most suitable patients of the survive class for the learning stage.

According to the previous analysis, we select EUSBoost and EUnderBagging as control methods to conduct the statistical study. That is, for each family we compare the control method (R+) versus the remainder methods (R−). The statistical results are reported in Tables 7 and 8 for boosting-based and bagging-based ensembles, respectively. From these results we can observe the following facts:

- EUSBoost is statistically outperforming AdaBoost.
- SMOTEBoost is clearly enhanced by EUSBoost in terms of AUC and GM.
- There are not statistical differences among EUSBoost and RUSBoost though the former provides better

TABLE 7. Results of the Mann-Whitney's U statistical test to compare EUSBoost (R+) versus the remainder boosting-based ensembles (R−).

Comparison	AUC			GM			F-Measure		
	p-value	(R+)	(R-)	p-value	(R+)	(R-)	p-value	(R+)	(R-)
EUSBoost vs. AdaBoost	$< 0.01^*$	71	$< 0.01^*$	130.42	70.58	$< 0.01^*$	113.64	87.36	
EUSBoost vs. RUSBoost	0.16	106.29	94.71	0.18	106.01	94.99	0.96	100.32	100.68
EUSBoost vs. SMOTEBoost	$< 0.01^*$	118.15	82.85	$< 0.01^*$	119.41	81.59	0.61	102.59	98.41

TABLE 8. Results of the Mann-Whitney's U statistical test to compare EUnderBagging (R+) versus the remainder bagging-based ensembles (R−).

Comparison	AUC			GM			F-Measure		
	p-value	(R+)	(R-)	p-value	(R+)	(R-)	p-value	(R+)	(R-)
EUnderBagging vs. OverBagging	$< 0.01^*$	118.95	82.05	$< 0.01^*$	118.75	82.25	$< 0.01^*$	111.77	89.23
EUnderBagging vs. SMOTEBagging	$< 0.01^*$	112.03	88.97	$< 0.01^*$	11.92	89.08	0.06^*	108.30	92.70
EUnderBagging vs. UnderBagging	0.34	104.38	96.62	0.31	104.69	96.31	0.36	104.23	96.77
EUnderBagging vs. UnderOverBagging	$< 0.01^*$	121.44	79.56	$< 0.01^*$	121.34	79.66	0.02^*	110.14	90.86

TABLE 9. Testing results obtained for EUnderBagging, EUSBoost and EasyEnsemble.

Classifier	AUC	GM	F-measure	TN_{rate}	TP_{rate}
EUSBoost	0.8293 ± 0.0721	0.8255 ± 0.0731	0.6509 ± 0.0960	0.7972 ± 0.0647	0.8614 ± 0.1283
EUnderBagging	**0.8498 ± 0.0648**	**0.8472 ± 0.0650**	**0.6838 ± 0.0912**	**0.8196 ± 0.0601**	**0.8800 ± 0.1090**
EasyEnseble	0.8294 ± 0.0651	0.8263 ± 0.0652	0.6522 ± 0.0899	0.7958 ± 0.0687	0.8631 ± 0.1091

TABLE 10. Results of the Mann-Whitney's U statistical test to compare EUnderBagging (R+) versus EUSBoost and EasyEnsemble (R−).

Comparison	AUC			GM			F-Measure		
	p-value	(R+)	(R-)	p-value	(R+)	(R-)	p-value	(R+)	(R-)
EUnderBagging vs. EUSBoost	0.04^*	108.9	92.1	0.03^*	109.29	91.71	0.01^*	110.83	90.17
EUnderBagging vs. EasyEnsemble	0.02^*	110.40	90.60	0.01^*	110.78	90.22	$< 0.01^*$	111.96	89.04

performance results and the p-values are low in terms of AUC and GM.

- EUnderBagging outperforms bagging-based ensembles using over sampling techniques (OverBagging, SMOTEBagging and UnderOverBagging).
- EUnderBagging achieves better results and rankings (in the statistical study) than UnderBagging in the three metrics but there are not statistical differences between them.

3) DETERMINING THE BEST ENSEMBLE METHOD TO TACKLE THIS PROBLEM

Taking into account both the performance results and the statistical analysis, we can select EUSBoost and EUnderBagging as the best options from each family of ensembles. Therefore, we compare them along with EasyEnsemble, as representative of hybrid ensemble approaches combining bagging and boosting, so that the best ensemble with sampling technique can be determined. Their performance results and the corresponding statistical comparison are shown in Tables 9 and 10, respectively. From these results we can conclude that EUnderBagging is clearly the best option since it provides the best performance in all the metrics and there are statistical differences versus EUSBoost and EasyEnsemble.

4) COMPARISON VERSUS CLASSICAL CLASSIFIERS TO DEAL WITH TRAUMA PATIENTS

Finally, we want to study if our new proposal, EUnderBagging, is able to enhance the performance of the models that are currently used by the Hospital of Navarre staff (MPMN and TRISS) as well as with $C45\_CS$, TRISS_{Nav} and MCS. The results of these six classifiers are reported in Table 11, where it can be observed that EUnderBagging improves the results of all of the remainder approaches. If we analyze in detail these results we can observe the following facts:

- The adaptation of the parameters of TRISS to the features of the patients in the MTRN (TRISS_{Nav}) allows to enhance the results of TRISS in all the metrics.
- Among the results of models composed of a unique classifier the ones provided by $C45\_CS$ are better in terms of AUC and GM as a consequence of a great improvement of the TP_{rate} at the cost of a large FP_{rate} $(1 - TN_{rate})$.
- Comparing the logistic regression-based methods with EUnderBagging we can find that the latter is a better choice as the results in all the balanced metrics are better and the FP_{rate} is not large. Obviously, the intepretability of the logistic regression models is lost when applying EUnderBagging.

TABLE 11. Testing results obtained for EUnderBagging and standard methods.

Classifier	AUC	GM	F-measure	TN_{rate}	TP_{rate}
TRISS	0.6964 ± 0.0769	0.6399 ± 0.1165	0.5300 ± 0.1446	0.9457 ± 0.0383	0.4470 ± 0.1484
TRISS$_{Nav}$	0.7093 ± 0.0808	0.6581 ± 0.1179	0.5538 ± 0.1514	0.9473 ± 0.0029	0.4713 ± 0.0082
MPMN	0.7632 ± 0.0768	0.7291 ± 0.1030	0.6517 ± 0.1300	**0.9631 ± 0.0348**	0.5633 ± 0.1526
$C45\_CS$	0.7780 ± 0.0675	0.7690 ± 0.0684	0.5719 ± 0.0844	0.7110 ± 0.0979	0.8449 ± 0.1343
MCS	0.7805 ± 0.0832	0.7661 ± 0.0961	0.6506 ± 0.1271	0.8908 ± 0.0135	0.6703 ± 0.0205
EUnderBagging	**0.8498 ± 0.0648**	**0.8472 ± 0.0650**	**0.6838 ± 0.0912**	0.8196 ± 0.0601	**0.8800 ± 0.1090**

TABLE 12. Results of the Mann-Whitney's U statistical test to compare EUnderBagging (R+) versus the remainder approaches (R−).

Comparison	AUC			GM			F-Measure		
	p-value	(R+)	(R-)	p-value	(R+)	(R-)	p-value	(R+)	(R-)
EUnderBagging vs. TRISS	$< 0.01^*$	143.84	57.16	$< 0.01^*$	145.53	55.47	$< 0.01^*$	132.02	68.98
EUnderBagging vs. TRISS$_{Nav}$	$< 0.01^*$	141.59	59.41	$< 0.01^*$	143.40	57.60	$< 0.01^*$	125.91	75.09
EUnderBagging vs. MPMN	$< 0.01^*$	131.30	69.70	$< 0.01^*$	133.95	67.05	0.07^*	107.93	93.07
EUnderBagging vs. $C45\_CS$	$< 0.01^*$	128.86	72.14	$< 0.01^*$	130.45	70.55	$< 0.01^*$	132.28	68.72
EUnderBagging vs. MCS	$< 0.01^*$	125.78	75.22	$< 0.01^*$	126.81	74.19	0.06^*	108.30	92.70

- The results of EUnderBagging are superior than those of MCS as the balance among TP_{rate} and TN_{rate} is better as well. We may guess that this behavior is caused by the effectiveness of the sampling method used by EUnderBagging.
- EUnderBagging is providing the most stable results as its standard deviation in terms of AUC, GM and F-measure is always less than the remainder methods (except C45_CS in the F-measure). Consequently, EUnderBagging is not only the best performing method but the most stable one.

In order to give statistical support to the results shown in Table 11, we have conducted the proper statistical study, whose results are shown in Table 12. These statistical results allow us to conclude that the usage of our new proposal is suitable to deal with the survival prediction of trauma patients.

VI. CONCLUSIONS

The prediction of the survival status of severe trauma patients is an important problem for emergency services at the hospitals. It is an imbalanced problem because the number of patients who survive to their injuries excels that of those who die. To solve these types of problems, the classical approaches used by doctors are based on logistic regression models. Such classifiers provide good results but, as every data mining method, they can suffer from the problems derived by the special features of imbalanced classification problems.

In this work we have proposed a new method, named EUnderBagging, to construct bagging-based ensembles for imbalanced problems by including an evolutionary under sampling process to obtain each bag. Consequently, EUnderBagging is specifically designed to tackle imbalanced classification problems. From the experimental study we can stress the following facts: 1) the combination of ensembles and sampling techniques is appropriate for this problem; 2) the usage of under sampling techniques is providing better results than over sampling techniques when combined with ensembles; 3) bagging-based ensembles work better than boosting-based ones; 4) EUnderBagging is the best choice among the considered ensembles that use sampling techniques and 5) our new proposal allows one to clearly enhance the behavior of the currently used methods by the staff of the Hospital of Navarre. New prediction systems are continuously being developed and consequently, we will test the behavior of other ensembles of classifiers like [60]–[63] in the future.

REFERENCES

[1] W. Haddon, Jr., "Advances in the epidemiology of injuries as a basis for public policy," *Public Health Rep.*, vol. 95, no. 5, pp. 411–421, 1980.

[2] C. Pape-Kohler, C. Simanski, U. Nienaber, and R. Lefering, "External factors and the incidence of severe trauma: Time, date, season and moon," *Injury*, vol. 45, pp. S93–S99, Oct. 2014.

[3] C. R. Boyd, M. A. Tolson, and W. S. Copes, "Evaluating trauma care: The TRISS method. Trauma score and the injury severity score," *J. Trauma*, vol. 27, no. 4, pp. 370–378, 1987.

[4] R. Lefering, S. Huber-Wagner, U. Nienaber, M. Maegele, and B. Bouillon, "Update of the trauma risk adjustment model of the TraumaRegister DGU: The Revised Injury Severity Classification, version II," *Crit. Care*, vol. 18, no. 5, p. 476, 2014.

[5] T. Belzunegui, C. Gradín, M. Fortún, A. Cabodevilla, A. Barbachano, and J. A. Sanz, "Major trauma registry of navarre (spain): The accuracy of different survival prediction models," *Amer. J. Emergency Med.*, vol. 31, no. 9, pp. 1382–1388, 2013.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Hoboken, NJ, USA: Wiley, 2001.

[7] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.

[8] P. Siriyasatien, S. Chadsuthi, K. Jampachaisri, and K. Kesorn, "Dengue epidemics prediction: A survey of the state-of-the-art based on data science processes," *IEEE Access*, vol. 6, pp. 53757–53795, 2018.

[9] J. A. Sanz, D. Bernardo, F. Herrera, H. Bustince, and H. Hagras, "A compact evolutionary interval-valued fuzzy rule-based classification system for the modeling and prediction of real-world financial applications with imbalanced data," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 4, pp. 973–990, Aug. 2015.

[10] J. A. Sanz, M. Galar, A. Jurio, A. Brugos, M. Pagola, and H. Bustince, "Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system," *Appl. Soft Comput.*, vol. 20, pp. 103–111, Jul. 2013.

[11] N. V. Chawla, N. Japkowicz, and A. Kolcz, "Editorial: Special issue on learning from imbalanced data sets," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 1–6, Jun. 2004.

[12] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern., C (Appl. Rev.)*, vol. 42, no. 4, pp. 463–484, Jul. 2012.

[13] Q. Yang and X. Wu, "10 Challenging problems in data mining research," *Int. J. Inf. Technol. Decis. Making*, vol. 5, no. 4, pp. 597–604, 2006.

[14] J. R. Quinlan, "Improved estimates for the accuracy of small disjuncts," *Mach. Learn.*, vol. 6, no. 1, pp. 93–98, 1991.

[15] G. Wu and E. Y. Chang, "KBA: Kernel boundary alignment considering imbalanced data distribution," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 786–795, Jun. 2005.

[16] G. E. Batista, R. C. Prati, and M. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newslett.*, vol. 6, no. 1, pp. 20–29, 2004.

[17] N. V. Chawla, D. A. Cieslak, L. O. Hall, and A. Joshi, "Automatically countering imbalance and its empirical relationship to cost," *Data Mining Knowl. Discovery*, vol. 17, no. 2, pp. 225–252, 2008.

[18] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 40, no. 1, pp. 185–197, Jan. 2010.

[19] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, Cybern., B (Cybern.)*, vol. 39, no. 2, pp. 539–550, Apr. 2009.

[20] S. Huda, J. Yearwood, H. F. Jelinek, M. M. Hassan, G. Fortino, and M. Buckland, "A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis," *IEEE Access*, vol. 4, pp. 9145–9154, 2016.

[21] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers, 1993.

[22] S. García and F. Herrera, "Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy," *Evol. Comput.*, vol. 17, no. 3, pp. 275–306, 2009.

[23] M. Galar, A. Fernández, E. Barrenechea, and F. Herrera, "EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling," *Pattern Recognit.*, vol. 46, no. 12, pp. 3460–3471, Dec. 2013.

[24] (2010). *Boletin Oficial de Navarra Numero 79 de 30 de Junio de 2010-navarra.es. (Consultado 10 mayo 2010)*. Accessed: Oct. 15, 2010. [Online]. Available: http://www.navarra.es/home_es/Actualidad/BON/Boletines/2010/79/Anuncio-16/

[25] K. M. Ting, "An instance-weighting method to induce cost-sensitive trees," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 3, pp. 659–665, May 2002.

[26] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997.

[27] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Mach. Learn.*, vol. 30, nos. 2–3, pp. 195–215, 1998.

[28] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. Newton, MA, USA: Butterworth-Heinemann, 1979.

[29] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, 1947.

[30] S. Polinder, W. Meerding, S. Mulder, E. Petridou, and E. van Beeck, "EUROCOST reference group. Assessing the burden of injury in six European countries," *Bull World Health Org.*, vol. 85, no. 1, pp. 27–34, 2007.

[31] D. A. Pollock and P. W. McClain, "Trauma registries: Current status and future prospects," *J. Amer. Med. Assoc.*, vol. 262, no. 16, pp. 2280–2283, 1989.

[32] K. G Ringdal, T. J. Coats, R. Lefering, S. Di Bartolomeo, P. A. Steen, O. Røise, L. Handolin, and H. M. Lossius, "The Utstein template for uniform reporting of data following major trauma: A joint revision by SCANTEM, TARN, DGU-TR and RITG," *Scand. J. Trauma, Resuscitation Emergency Med.*, vol. 16, no. 1, p. 7, 2008.

[33] S. P. Baker, B. O'Neill, W. Haddon, Jr., and W. Long, "The injury severity score: A method for describing patients with multiple injuries and evaluating emergency care," *J. Trauma*, vol. 14, no. 3, pp. 187–196, 1974.

[34] T. Osler, S. P. Baker, and W. Long, "A modification of the injury severity score that both improves accuracy and simplifies scoring," *J. Trauma*, vol. 43, no. 6, pp. 922–925, 1997.

[35] H. R. Champion, W. J. Sacco, W. S. Copes, D. Gann, T. A. Gennarelli, and M. E. Flanagan, "A revision of the trauma score," *J. Trauma*, vol. 29, no. 5, pp. 623–629, 1989.

[36] C. G. Purroy, T. B. Otano, B. B. Fraile, R. Teijeira, M. F. Moral, and D. R. Diez, "Changes in the characteristics and incidence of multiple-injury accidents in the Navarre community over a 10-year period," *Emergencias*, vol. 27, no. 3, pp. 174–180, 2015.

[37] W.-S. Chen, S.-W. Lee, S. Jamaluddin, and C.-P. Wong, "Comparison of trauma and injury severity score model with alternative approach in outcome prediction in trauma using national trauma database in malaysia," *Trauma*, vol. 19, no. 2, pp. 103–112, 2017.

[38] C. de Alencar Domingues, R. Coimbra, R. S. Poggetti, L. de Souza Nogueira, and R. C. Sousa, "Performance of new adjustments to the triss equation model in developed and developing countries," *World J. Emergency Surg.*, vol. 12, no. 1, p. 17, 2017.

[39] C. O. Valderrama-Molina, N. Giraldo, A. Constain, A. Puerta, C. Restrepo, A. León, and F. Jaimes, "Validation of trauma scales: Iss, niss, rts and triss for predicting mortality in a colombian population," *Eur. J. Orthopaedic Surg. Traumatol.*, vol. 27, no. 2, pp. 213–220, 2017.

[40] C. de Alencar Domingues, R. Coimbra, R. S. Poggetti, L. de Souza Nogueira, and R. M. C. de Sousa, "New trauma and injury severity score (TRISS) adjustments for survival prediction," *World J. Emergency Surg.*, vol. 13, no. 1, p. 12, 2018.

[41] N. O. Skaga, T. Eken, and S. Sovik, "Validating performance of TRISS, TARN and NORMIT survival prediction models in a Norwegian trauma population," *Acta Anaesthesiologica Scandinavica*, vol. 62, no. 2, pp. 253–266, 2018.

[42] J. A. Barea-Mendoza, M. Chico-Fernandez, M. Sanchez-Casado, I. Molina-Diaz, M. Quintana-Diaz, J. M. Jiménez-Moragas, J. Perez-Barcena, and J. A. Llompart-Pou, "Predicting survival in geriatric trauma patients: A comparison between the triss methodology and the geriatric trauma outcome score," *Cirugía Española*, vol. 96, no. 6, pp. 357–362, 2018.

[43] H. R. Champion, W. S. Copes, W. J. Sacco, M. M. Lawnick, S. L. Keast, L. W. Bain, Jr., M. E. Flanagan, and C. F. Frey, "The major trauma outcome study: Establishing national norms for trauma care," *J. Trauma-Injury, Infection Crit. Care*, vol. 30, no. 11, pp. 1356–1365, 1990.

[44] R. Lefering, "Development and validation of the revised injury severity classification score for severely injured patients," *Eur. J. Trauma Emergency Surg.*, vol. 35, no. 5, pp. 437–447, 2009.

[45] B. A. Ali, M. F. Moral, T. B. Otano, D. R. Diez, and M. C. Neira, *Escalas Para Predicción de Resultados Tras Traumatismo Grave*, 2017, pp. 103–118.

[46] B. Ali, R. Lefering, M. Moral, and T. Otano, "Validación del modelo de predicción de mortalidad deNavarra y comparación con el revised injury severityclassification score II en los pacientes con traumatismograve atendidos por el sistema de emergencias de navarra," *Emergencias*, vol. 30, no. 2, pp. 98–104, 2018.

[47] J. Sanz, J. Fernandez, H. Bustince, C. Gradin, M. Fortun, and T. Belzunegui, "A decision tree based approach with sampling techniques to predict the survival status of poly-trauma patients," *Int. J. Comput. Intell. Syst.*, vol. 10, no. 1, pp. 440–455, 2017.

[48] J. Sanz, D. Paternain, M. Galar, J. Fernandez, D. Reyero, and T. Belzunegui, "A new survival status prediction system for severe trauma patients based on a multiple classifier system," *Comput. Methods Programs Biomed.*, vol. 142, pp. 1–8, Apr. 2017.

[49] A. Orriols-Puig and E. Bernadó-Mansilla, "Evolutionary rule-based systems for imbalanced data sets," *Soft Comput.*, vol. 13, no. 3, pp. 213–225, 2009.

[50] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[51] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, *SMOTEBoost: Improving Prediction of the Minority Class in Boosting*, vol. 2838. Berlin, Germany: Springer, 2003, pp. 107–119.

[52] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[53] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[54] S. Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models," in *Proc. IEEE Symp. Comput. Intell. Data Mining*, Mar./Apr. 2009, pp. 324–331.

[55] R. Barandela, R. M. Valdovinos, and J. S. Sanchez, "New applications of ensembles of classifiers," *Pattern Anal. Appl.*, vol. 6, no. 3, pp. 245–256, Dec. 2003.

[56] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[57] L. J. Eshelman, "The CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination," in *Foundations of Genetic Algorithms*. Burlington, MA, USA: Morgan Kaufman, 1991, pp. 265–283.

[58] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, NJ, USA: Wiley, 2004.

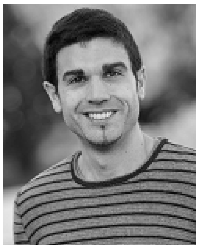
[59] G. U. Yule, "On the association of attributes in statistics: With illustrations from the material of the childhood society, &c," *Philos. Trans. Roy. Soc. London A, Math. Phys. Sci.*, vol. 194, no. 1900, pp. 257–319, 1900.

[60] B. Sun, H. Chen, J. Wang, and H. Xie, "Evolutionary under-sampling based bagging ensemble method for imbalanced data classification," *Frontiers Comput. Sci.*, vol. 12, no. 2, pp. 331–350, 2018.

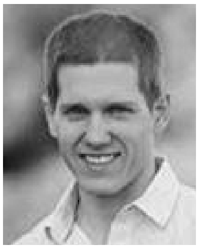
[61] M. A. Tahir, J. Kittler, and F. Yan, "Inverse random under sampling for class imbalance problem and its application to multi-label classification," *Pattern Recognit.*, vol. 45, no. 10, pp. 3738–3750, 2012.

[62] J. Gong and H. Kim, "RHSBoost: Improving classification performance in imbalance data," *Comput. Statist. Data Anal.*, vol. 111, pp. 1–13, Jul. 2017.

[63] Y. Qian, Y. Liang, M. Li, G. Feng, and X. Shi, "A resampling ensemble algorithm for classification of imbalance problems," *Neurocomputing*, vol. 143, pp. 57–67, Nov. 2014.



JOSE ANTONIO SANZ received the M.Sc. and Ph.D. degrees in computer science from the Universidad Publica de Navarra, Spain, in 2008 and 2011, respectively, where he is currently an Assistant Professor with the Department of Statistics, Computer Science, and Mathematics. He has authored 32 published original articles in international journals. He has published over 45 contributions in conferences. He is a member of the European Society for Fuzzy Logic and Technology (EUSFLAT) and the Spanish Association of Artificial Intelligence (AEPIA). His research interests include fuzzy techniques for classification problems, interval-valued fuzzy sets, genetic fuzzy systems, and medical applications using soft-computing techniques. He received the Best Paper Award from FLINS 2012 International Conference and the Pepe Millá Award, in 2014.



MIKEL GALAR (M'16) received the M.Sc. and Ph.D. degrees in computer science from the Universidad Publica de Navarra, Pamplona, Spain, in 2009 and 2012, respectively, where he is currently an Assistant Professor with the Department of Statistics, Computer Science, and Mathematics. He is the author of 31 published original articles in international journals and more than 45 contributions to conferences. He is also a Reviewer of more than 35 international journals. His research interests include data mining, classification, multi-classification, ensemble learning, evolutionary algorithms, fuzzy systems, and big data. He is a member of the European Society for Fuzzy Logic and Technology (EUSFLAT) and the Spanish Association of Artificial Intelligence (AEPIA). He has received the Extraordinary Prize for the Ph.D. thesis from the Public University of Navarre and the 2013 IEEE Transactions on Fuzzy System Outstanding Paper Award for the paper *A New Approach to Interval-Valued Choquet Integrals and the Problem of Ordering in Interval-Valued Fuzzy Set Applications*, in 2016.

HUMBERTO BUSTINCE (M'08–SM'15) received the bachelor's degree in physics from the University of Salamanca, in 1983, and the Ph.D. degree in mathematics from the Universidad Publica de Navarra, Pamplona, Spain, in 1994, where he is currently a Full Professor of computer science and artificial intelligence. He is a main Researcher of the Artificial Intelligence and Approximate Reasoning Group, whose main research lines are both theoretical (aggregation functions, information and comparison measures, fuzzy sets, and extensions) and applied (image processing, classification, machine learning, data mining, and big data). He has led 11 I + D public-funded research projects at a national and a regional level. He is currently the main Researcher of scientific network about fuzzy logic and soft computing and the Spanish Science Program Project. He has been the in charge of research projects collaborating with private companies. He has taken part in two international research projects. He has authored more than 210 works, according to Web of Science, in conferences and international journals, with around 110 of them in journals of the first quartile of JCR. Moreover, five of these works are also among the highly-cited papers of the last ten years, according to Science Essential Indicators of Web of Science. He is the coauthor of a monograph about averaging functions and the coeditor of several books. He is the Editor-in-Chief of the online magazine *Mathware & Soft Computing* of the European Society for Fuzzy Logic and Technologies and of the *Axioms* journal. He is an Associate Editor of the IEEE Transactions on Fuzzy Systems Journal and a Member of the Editorial Board of the journals such as *Fuzzy Sets and Systems*, *Information Fusion*, the *International Journal of Computational Intelligence Systems*, and the *Journal of Intelligent & Fuzzy Systems*. He has organized some renowned international conferences such as EUROFUSE 2009 and AGOP 2013. He is a Fellow of the International Fuzzy Systems Association.

TOMAS BELZUNEGUI received the bachelor's and Ph.D. degrees in medicine from the University of Navarra, Pamplona, Spain, in 1979 and 1991, respectively. He is Emergency Physician of the Navarre Health Service. He is currently the Director of Emergencies and Hospitalization, Hospital of Navarra (1,100-bed Public Center located in Pamplona). He has been responsible for the strategy of attention to the time-depending pathologies on the health plan with the Health Department of Navarra, since 2014. He is currently an Associate Professor of the Department of Health, Public University of Navarre. He has led 17 I + D public-funded research projects at national and regional level. He is the author of 73 published original articles in international journals and over 120 contributions to conferences. He is the President of the Research Commission of the Navarra Health Service and he is also the President of the Teaching and Research Committee of the Hospital of Navarra.