



Comparison of Resampling Techniques for Imbalanced Datasets in Machine Learning: Application to Epileptogenic Zone Localization From Interictal Intracranial EEG Recordings in Patients With Focal Epilepsy

Giulia Varotto^{1,2*}, Gianluca Susi^{3,4}, Laura Tassi⁵, Francesca Gozzo⁵, Silvana Franceschetti² and Ferruccio Panzica⁶

¹ Epilepsy Unit, Bioengineering Group, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy, ² Neurophysiopathology Unit, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy, ³ Universidad Complutense de Madrid-Universidad Politécnica de Madrid (UPM-UCM) Laboratory of Cognitive and Computational Neuroscience, Center of Biomedical Technology, Technical University of Madrid, Madrid, Spain, ⁴ Department of Experimental Psychology, Cognitive Processes and Logopedy, Complutense University of Madrid, Madrid, Spain, ⁵ “Claudio Munari” Epilepsy Surgery Centre, Niguarda Hospital, Milan, Italy, ⁶ Clinical Engineering, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy

OPEN ACCESS

Edited by:

Sharlene D. Newman,
University of Alabama, United States

Reviewed by:

Leon D. Iasemidis,
Louisiana Tech University,
United States
Joseph R. Madsen,
Boston Pediatric Neurosurgical
Foundation, United States

*Correspondence:

Giulia Varotto
giulia.varotto@istituto-besta.it
orcid.org/0000-0001-8849-0398

Received: 26 May 2021

Accepted: 23 September 2021

Published: 19 November 2021

Citation:

Varotto G, Susi G, Tassi L, Gozzo F, Franceschetti S and Panzica F (2021) Comparison of Resampling Techniques for Imbalanced Datasets in Machine Learning: Application to Epileptogenic Zone Localization From Interictal Intracranial EEG Recordings in Patients With Focal Epilepsy. *Front. Neuroinform.* 15:715421. doi: 10.3389/fninf.2021.715421

Aim: In neuroscience research, data are quite often characterized by an imbalanced distribution between the majority and minority classes, an issue that can limit or even worsen the prediction performance of machine learning methods. Different resampling procedures have been developed to face this problem and a lot of work has been done in comparing their effectiveness in different scenarios. Notably, the robustness of such techniques has been tested among a wide variety of different datasets, without considering the performance of each specific dataset. In this study, we compare the performances of different resampling procedures for the imbalanced domain in stereo-electroencephalography (SEEG) recordings of the patients with focal epilepsies who underwent surgery.

Methods: We considered data obtained by network analysis of interictal SEEG recorded from 10 patients with drug-resistant focal epilepsies, for a supervised classification problem aimed at distinguishing between the epileptogenic and non-epileptogenic brain regions in interictal conditions. We investigated the effectiveness of five oversampling and five undersampling procedures, using 10 different machine learning classifiers. Moreover, six specific ensemble methods for the imbalanced domain were also tested. To compare the performances, Area under the ROC curve (AUC), F-measure, Geometric Mean, and Balanced Accuracy were considered.

Results: Both the resampling procedures showed improved performances with respect to the original dataset. The oversampling procedure was found to be more sensitive to the type of classification method employed, with Adaptive Synthetic Sampling

(ADASYN) exhibiting the best performances. All the undersampling approaches were more robust than the oversampling among the different classifiers, with Random Undersampling (RUS) exhibiting the best performance despite being the simplest and most basic classification method.

Conclusions: The application of machine learning techniques that take into consideration the balance of features by resampling is beneficial and leads to more accurate localization of the epileptogenic zone from interictal periods. In addition, our results highlight the importance of the type of classification method that must be used together with the resampling to maximize the benefit to the outcome.

Keywords: imbalanced dataset classification, re-sampling techniques, oversampling and undersampling, ensemble methods, network analysis, epilepsy surgery, stereo-EEG/intracranial recordings, epileptogenic zone localization

INTRODUCTION

Epilepsy is a chronic neurological disease affecting 1% of the worldwide population (Fiest et al., 2017). Approximately 30% of the patients with focal epilepsies are resistant to the antiepileptic drugs (AEDs), and they can be considered as candidate for epilepsy surgery, with the aim of removing the epileptogenic zone (EZ). The latter is defined as the minimum amount of cortex that must be resected (inactivated or completely disconnected) to produce seizure freedom (Lüders et al., 2006; Ryvlin et al., 2014). However, the correct localization of the EZ to achieve seizure freedom after surgery, is still an unsolved and open question, as indicated by the high rate of failure of seizure control (30–40%) after surgery (Spencer and Huh, 2008; Bulacio et al., 2012). The advanced signal processing approaches, especially those based on the connectivity analysis, have been largely applied to stereo-electroencephalography (SEEG) from the patients with epilepsy to better pinpoint the location of the EZ (Varotto et al., 2013; Bartolomei et al., 2017; Adkinson et al., 2019; Narasimhan et al., 2020).

The supervised machine learning methods are increasingly applied in epilepsy research, representing useful tools to integrate the complex and large-scale data deriving from different electrophysiological or imaging techniques, such as EEG, magnetoencephalography (MEG), functional-MRI (fMRI), or positron emission tomography (PET) (refer to Abbasi and Goldenholz, 2019 for a comprehensive review). Most of these studies focused on the following aspects: diagnosis of epilepsy (Kassahun et al., 2014; Azami et al., 2016; Soriano et al., 2017), seizure prediction (Acharya et al., 2018; Kiral-Kornek et al., 2018; Daoud and Bayoumi, 2019), lateralization of temporal lobe epilepsy (Jin and Chung, 2017; Frank et al., 2018; Peter et al., 2018), and post-surgical outcome prediction (Armañanzas et al., 2013; Goldenholz et al., 2016; Gleichgerrcht et al., 2018). With respect to the localization of the EZ and support to pre-surgical planning, few works applied machine learning tools, showing the promising usefulness of this approach, and the need for further investigation and generalization (Dian et al., 2015; Elahian et al., 2017; Khambhati et al., 2017; Roland et al., 2017). In this specific framework, one central issue that should be taken into account, and which could represent one of the main limitations, is that

the EZ represents a smaller region compared with the other non-EZ areas explored. This leads to an uneven distribution of the majority (non-EZ) and minority (EZ) classes, which can strongly worsen or limit the classification performances. This situation is known as the class imbalance problem and can be considered one of the central topics in machine learning research (He and Garcia, 2009; Ali et al., 2015; Fernández et al., 2018).

In the past decade, many different approaches have been developed to cope with imbalanced classification, most of them based on four different families: resampling techniques, cost-sensitive learning, algorithm modification, and ensemble methods (Mena and Gonzalez, 2006; Galar et al., 2012; Krawczyk et al., 2014; Loyola-González et al., 2016).

Among these, the methods belonging to the data resampling family have been proved useful as well as relatively simple approaches to be applied in the medical context (Lee, 2014; Loyola-González et al., 2016). In data resampling, the training instances are modified to rebalance the class distribution through *oversampling* of the minority class, or *undersampling* of the majority one, before training the classifier. Oversampling could have the limitation of overfitting the minority class, while undersampling could eliminate potential useful information for correct classification (Chawla, 2009).

Different studies dealt with the comparisons of performances of most of the existing resampling techniques, most of which were applied to a wide variety of datasets together, being mainly aimed at assessing the robustness of results across different dataset combinations (López et al., 2013). Nevertheless, when applied to a single specific dataset, such comparison can lead to different results (Xie et al., 2020), reflecting a lack of consensus about the performances of such techniques and putting in evidence the need for *ad-hoc* comparisons in each specific clinical framework.

To the best of our knowledge, this is the first study focused on the evaluation and comparison of these approaches in the context of epilepsy, and in particular, in the framework of the surgical planning based on analysis of electrophysiological intracranial recordings.

In this study, we compared five oversampling and five undersampling procedures and tested the resulting rebalanced datasets with 10 different machine learning classifiers (such as

both standard machines and classical ensemble approaches). Moreover, six specific ensemble methods properly modified for imbalanced domain and belonging to data variation-based ensemble were tested and compared. In these algorithms, the resampling phase is applied to each step of the ensemble classifier, in such a way that each classifier is trained with a different resampled dataset (Galar et al., 2012). For this reason, we considered them as an extension of resampling methods, which need to be compared with the oversampling and undersampling techniques combined with the classical ensemble approaches.

The classification was based on the features obtained by network analysis of interictal SEEG recorded from the 10 patients who underwent epilepsy surgery and were seizure-free (SF) after 3 years of follow-up.

To compare the performances, *area under the ROC curve* (AUC), *balanced accuracy* (BalACC), *F-measure* (Fm), and *geometric mean* (Gmean) were used as metrics, since these are usually considered suitable measures to deal with the imbalanced datasets (Bekkar et al., 2013; López et al., 2013).

MATERIALS AND METHODS

We start this section by describing the steps of selection and signal recording of the patients. The methodological pipeline is then outlined: feature extraction, data resampling, classification, and evaluation of the performance of the model (as shown in **Figure 1** for a schematic representation). Finally, we describe the statistical analysis, which has been performed to evaluate the consistency of our results.

Selection of Patients

The study involved SEEG signals recorded from $N_p = 10$ patients (three women) with drug-resistant focal epilepsy at the Claudio Munari Epilepsy Surgery Center of Niguarda Hospital (Milan, Italy). The patients were selected from the 41 patients implanted with SEEG electrodes over 24 months. Among them, 24 had negative MRI and 10 of them were seizure-free after at least 3 years of follow-up and were finally considered for this study. **Table 1** presents the details of the main clinical features.

The mean age of the patients was 31.7 ± 7.3 years, and the mean duration of epilepsy was 17.2 ± 7.8 years. They had no obvious risk factor for epilepsy. The surgical outcome was assessed after at least 3 years of follow-up after surgery (mean follow-up period: 56 ± 13 months) and classified as class I according to Engel's classification (Engel, 1993).

SEEG Recordings

Stereo-electroencephalography signals were recorded using the multi-lead platinum-iridium electrodes (Dixi, Besançon, France, with 5–18 contacts of diameter 0.8 mm; 1.5 mm long; and 2 mm apart), implanted under general anesthesia after stereo-arteriography using a 3D MRI imported into a computer-assisted neuronavigational module to localize the blood vessels and guide electrode trajectory. The placement of intracerebral electrodes was defined according to the data derived by non-invasive anatomic-electroclinical procedures (Talairach and Bancaud, 1966; Cardinale et al., 2019).

The SEEG signals were recorded using a common reference electrode (Nikon-Kohden system; 192-channels; sampling rate 1 kHz) under video and clinical control over 5–20 days and then examined by the two expert neurologists to define the EZ and plan the surgical approach and resection. EZ was defined by considering ictal discharge recordings, responses associated with the intracerebral electrical stimulations, and neurophysiological mapping, and then integrated into the definition of the brain area(s) to be surgically excised. Post-resection MRI was used to identify the areas of the brain that were effectively removed. The target value to assess the classification performances—SEEG leads as belonging to EZ or non-EZ—was defined by considering the intersection between the group of SEEG leads labeled as EZ by the clinicians through the pre-surgical evaluation, and the resected zone.

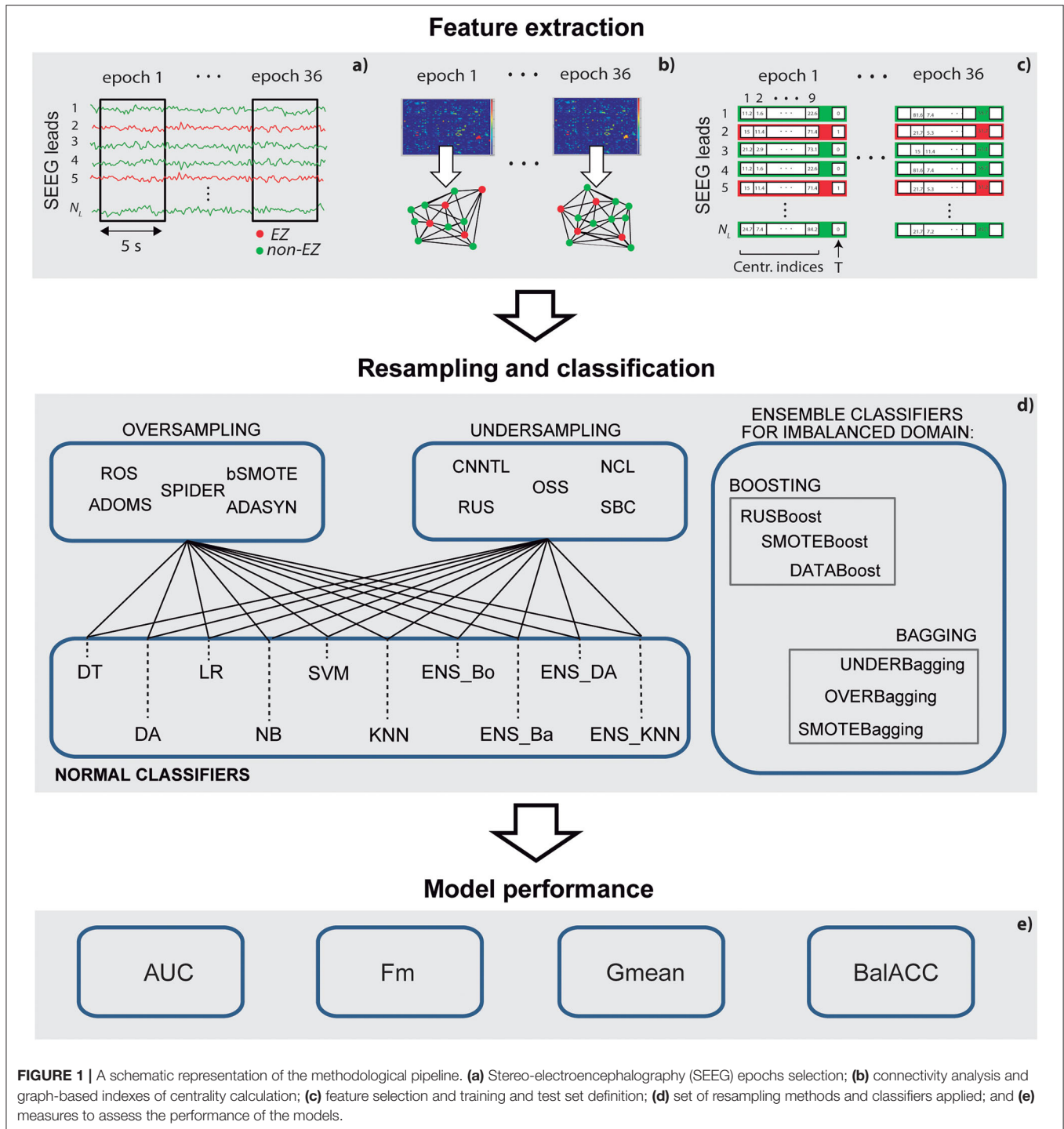
Feature Extraction

Stereo-electroencephalography signals were analyzed using bipolar derivations, and those presenting non-physiological artifacts were excluded from the analysis. The number of analyzed SEEG leads differed for each patient being on average $N_L = 73 \pm 6$. Furthermore, 3 min of continuous interictal SEEG signals, recorded during awake condition at least 1 h far from any ictal event, were selected and divided into $N_E = 36$, five s length, non-overlapping epochs. After testing several lengths and epochs partitions, 3 min length was selected as the minimum recording time to obtain a good EZ classification. The broad 1–80 Hz frequency band was used for the analysis. In addition, 36 time-varying connectivity matrices were estimated by applying a bivariate non-linear method and the non-linear regression index (h_2) (Lopes da Silva et al., 1989; Wendling et al., 2010) (refer to **Supplementary Material**). In this regard, a wide variety of methods have been proposed to estimate the SEEG connectivity, all of them being characterized by different advantages and pitfalls strongly depending on the signal and the aim of the study (Silfverhuth et al., 2012; Olejarczyk et al., 2017). Among them, a non-linear regression analysis has been proved to be particularly suitable to estimate the connectivity from the simulated coupled neuronal population (Wendling et al., 2009), and has been largely applied in the specific context of intracranial EEG recordings and EZ localization (Bartolomei et al., 2017).

From the adjacency matrices, the corresponding graphs were built for each patient, after applying a threshold to select the minimum number of connections that ensures a connected graph for all the epochs.

After a preliminary analysis involving several graph theory-based indices, nine of them, focusing on different complementary network properties of centrality (Oldham et al., 2019), were identified as the optimal one to classify EZ in the whole group of patients, and used as features of the classifier: *outdegree centrality* (C_e), *indegree* C_e , *outstrength* C_e , *instrength* C_e , *betweenness* C_e , *outcloseness* C_e , *incloseness* C_e , *pagerank* C_e , and *eigenvector* C_e . (as shown in **Supplementary Material** for a detailed description of the basic properties of these metrics).

The connectivity analysis was performed through a specific custom-written toolbox developed in Matlab (R20a; MathWorks Inc., Natick, MA, USA). Matlab graph toolbox and the Brain



connectivity toolbox (Rubinov and Sporns, 2010), were used for graph analysis.

To provide the classifiers with a suitable number of trials, we first grouped all the values of the features pertaining to the different time epochs and obtained, for each patient p , a matrix with $N_{L,p} \times N_E$ rows and 10 columns (i.e., nine features and one target). EZ has been considered as the positive class, with 1

indicating the EZ class and 0 the non-EZ class. The imbalanced ratio (IR)—the ratio between the number of trials pertaining to positive and negative classes—for each patient, is indicated in **Table 2**.

Since one of the main objectives of the proposed procedure was to classify SEEG signals of every single patient independently from the others, training and test set were defined by considering

TABLE 1 | Main clinical features, epileptogenic zone (EZ) localization performed by the standard methods, and surgery outcomes for the patients enrolled in this study.

Id	Gender	Age	Onset	Sz/m	Side	Lobe	Histology	Follow-Up/M	AEDs
1	M	27	19	10	R	TFI	crypto	68	Reduced
2	M	25	4	10	L	TI	crypto	48	Stopped
3	F	30	16	5	R	T	crypto	54	Reduced
4	F	27	17	10	R	T	crypto	46	Stopped
5	M	40	16	30	L	F	no	70	Reduced
6	M	39	20	1	R	F	crypto	42	Reduced
7	M	28	11	3	L	F	crypto	65	Ongoing
8	M	22	16	15	L	TO	crypto	34	Reduced
9	M	44	22	10	R	TO	FCD Ib	62	Reduced
10	F	35	4	5	R	TPCF	FCD Ia	71	Ongoing

AEDs, antiepileptic drugs; crypto, cryptogenic; FCD, focal cortical dysplasia; F, female; FC, fronto-central; FCD, focal cortical dysplasia; Fr, Frontal; HS, hippocampal sclerosis; M, male; Sz/m, seizures per month; Age, age at surgical intervention; Onset, age of epilepsy onset; PCI, parieto-centro-insular; T, Temporal; TFI, temporo-fronto-insular; TI, temporo insular; TC, temporo-central; TO, temporo-occipital; TPCF, temporo-parieto-centro-frontal. AEDs column refers to variation of drug therapy with respect to pre-surgical condition.

TABLE 2 | Number of analyzed SEEG leads (Total SEEG leads), number of leads belonging to the EZ (EZ leads), and Imbalanced ratio (IR) per patient.

Pt id	Tot SEEG leads	EZ leads	IR
1	66	5	12.2
2	73	6	11.2
3	80	7	10.4
4	81	9	8.0
5	62	4	14.5
6	72	2	35.0
7	76	8	8.5
8	78	13	5.0
9	72	5	13.4
10	72	7	9.3

a proportion of 9:1, using features from nine patients for training and features from one single patient for test. For further statistical analysis, the same splitting was repeated for all the combinations of patients, thus providing 10 different training-testing datasets.

Data Resampling

In all the patients, more electrode contacts were implanted in the non-epileptogenic than epileptogenic regions. This fact is reflected in a smaller number of EZ trials than the non-EZ trials, giving rise to the problems with the statistics of the applied classification methods (and hence, the subsequent learning by machine learning models).

Among the existing resampling techniques to tackle such class imbalance problems, we selected five methods of oversampling and five methods of undersampling and compared the performance of classifiers with respect to the original dataset.

The oversampling methods are based on the creation of a new bigger dataset, obtained by replicating or creating new samples, usually from the minority class:

- *Adaptive Synthetic Sampling* (ADASYN). ADASYN generates data considering a weighted distribution for different minority

class examples, where more synthetic data are generated for minority class examples that are harder to learn compared with those easier to learn (He et al., 2008).

- *Adjusting the direction of the synthetic minority class example* (ADOMS). ADOMS generates positive data instances from other instances in the original dataset selecting k as the nearest neighbors and using them to perform arithmetical operations to generate the new instance by principal component analysis (PCA) (Tang and Chen, 2008).
- *Random oversampling* (ROS). ROS generates minority class instances randomly (Batista et al., 2004).
- *Selective Pre-processing for Imbalanced Data* (SPIDER). SPIDER oversamples instances from the minority class that are difficult to learn and, at the same time, filters the examples from the majority class which are also difficult to learn (Stefanowski and Wilk, 2008).
- *Borderline-Synthetic Monitoring Oversampling Technique* (bSMOTE). The bSMOTE generates positive data instances from other instances in the original dataset selecting k as the nearest neighbors and using them to perform the arithmetical operations to generate the new instance (Han et al., 2005).

The undersampling methods are based on the reduction of the original dataset by eliminating samples, usually from the majority class:

- *Condensed Nearest Neighbor + Tomek’s modification of Condensed Nearest Neighbor* (CNNTL). CNNTL applies the CNN method and the Tomek Links method in a chain to delete the instances that lead us to misclassify new instances in the imbalanced domains (Batista et al., 2004).
- *Neighborhood Cleaning Rule* (NCL). NCL finds a subset S of the training set T applying the neighborhood cleaning rule of examples (Laurikkala, 2001).
- *One Side Selection* (OSS). OSS finds a subset S of the training set T applying the OSS of examples (Kubat and Matwin, 1997).
- *Random Undersampling* (RUS). RUS deletes the majority of class data instances randomly (Batista et al., 2004).

- *Undersampling based on clustering* (SBC). After dividing all the training samples into some clusters, SBC selects a suitable number of majority class samples from each cluster by considering the ratio of the number of majority class samples to the number of minority class samples in the cluster (Yen and Lee, 2006).

For both oversampling and undersampling methods, the default parameters were used. The corresponding parameters set can be found in the method library of KEEL software (UGR Granada, Spain) (Alcalá-Fdez et al., 2011).

Classification

To classify and compare the different resampled datasets, 10 different machine learning algorithms, belonging to the family of supervised classification, and most used in the contest of neurophysiological signal processing, were applied as follows:

1. Decision tree (DT): coarse tree, whose maximum number of branch points is set to 4. The method adopts the Gini's diversity index as the split criterion and envisages a pruning procedure.
2. Discriminant analysis (DA): creates non-linear boundaries between the classes (quadratic discriminant analysis).
3. Logistic regression (LR).
4. Naïve Bayes (NB): the method supports continuous attributes by assuming a Gaussian distribution (Gaussian Naïve Bayes).
5. Support vector machine (SVM): characterized by coarse distinctions between the classes, with kernel scale set to $4\sqrt{P}$, where P is the number of predictors (Coarse Gaussian SVM).
6. KNN (K-nearest neighbors): where we set the number of neighbors to 100 (Coarse distinctions between classes) and used the Euclidean distance metric (coarse KNN).
7. Boosted Ensemble (EnsBO): ensemble classifier which uses the meta-algorithm AdaBoost (Freund and Schapire, 1999).
8. Bagged Ensemble (EnsBA), Random forest Bag, with DT learners. This implementation uses Breiman's "random forest" algorithm (Breiman, 2001).
9. Discriminant Analysis Ensemble (EnsDA): combines different feature subsets to improve the classification performance (subspace ensemble), and uses Discriminant learners.
10. KNN ensemble (EnsKNN): Subspace ensemble with Nearest Neighbor learners.

During the training phase, the validation step was performed through a 5-fold cross-validation approach. For all the considered methods, default parameters were used. The corresponding parameters set can be found in the *Matlab classification learner* toolbox specification.

Ensemble Methods for Imbalanced Domain

Since the main objective of the study was to compare the effect of different resampling techniques on the classifier performances, in the previous section we described both the standard and classical ensemble classifiers, with the resampling procedure applied before the classification.

However, in the past years, ensemble-based classifiers have been considered a suitable approach in the imbalanced domain,

leading to the implementation of specific modification of the ensemble algorithm, in which the data rebalancing pre-processing is integrated into the ensemble algorithm and done before the learning stage of each classifier of the ensemble (Chawla et al., 2003; Seiffert et al., 2010). For this reason, we also tested six of these approaches, three belonging to boosting (methods 1–2–3) and three to bagging (methods 4–5–6) approach:

1. DATABoost: it combines the AdaBoost algorithm with a data generation strategy. It first identifies hard examples (seeds) and then carries out a rebalance process, always for both the classes (Guo and Viktor, 2004).
2. RUSBoost: multi-class AdaBoost with RUS in each iteration (Seiffert et al., 2010).
3. SMOTEBoost: multiclass AdaBoost with SMOTE in each operation (Chawla et al., 2003).
4. OVERBag: bagging with oversampling of the minority class (Wang and Yao, 2009).
5. SMOTEBag: bagging where SMOTE quantity of each bag varies (Wang and Yao, 2009).
6. UnderBag: bagging with undersampling of the majority class (Barandela et al., 2003b).

Performances Metrics

In common practice, accuracy is the most used measure to assess classifier performance. However, since it does not allow to distinguish between the number of correctly classified instances of the two different classes, it can lead to an erroneous conclusion when applied in the context of imbalanced datasets.

To assess and compare the performances of the classifiers, we used the following four metrics, which have been proven to be suitable for the imbalanced domain (Bekkar et al., 2013; López et al., 2013; Fernández et al., 2018):

$$AUC = \frac{1 + TPr + FPr}{2}$$

$$Fm = \frac{(1 + \beta^2)(PPV \cdot TPr)}{\beta^2 \cdot PPV + TPr}$$

$$GMean = \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{FP + TN}}$$

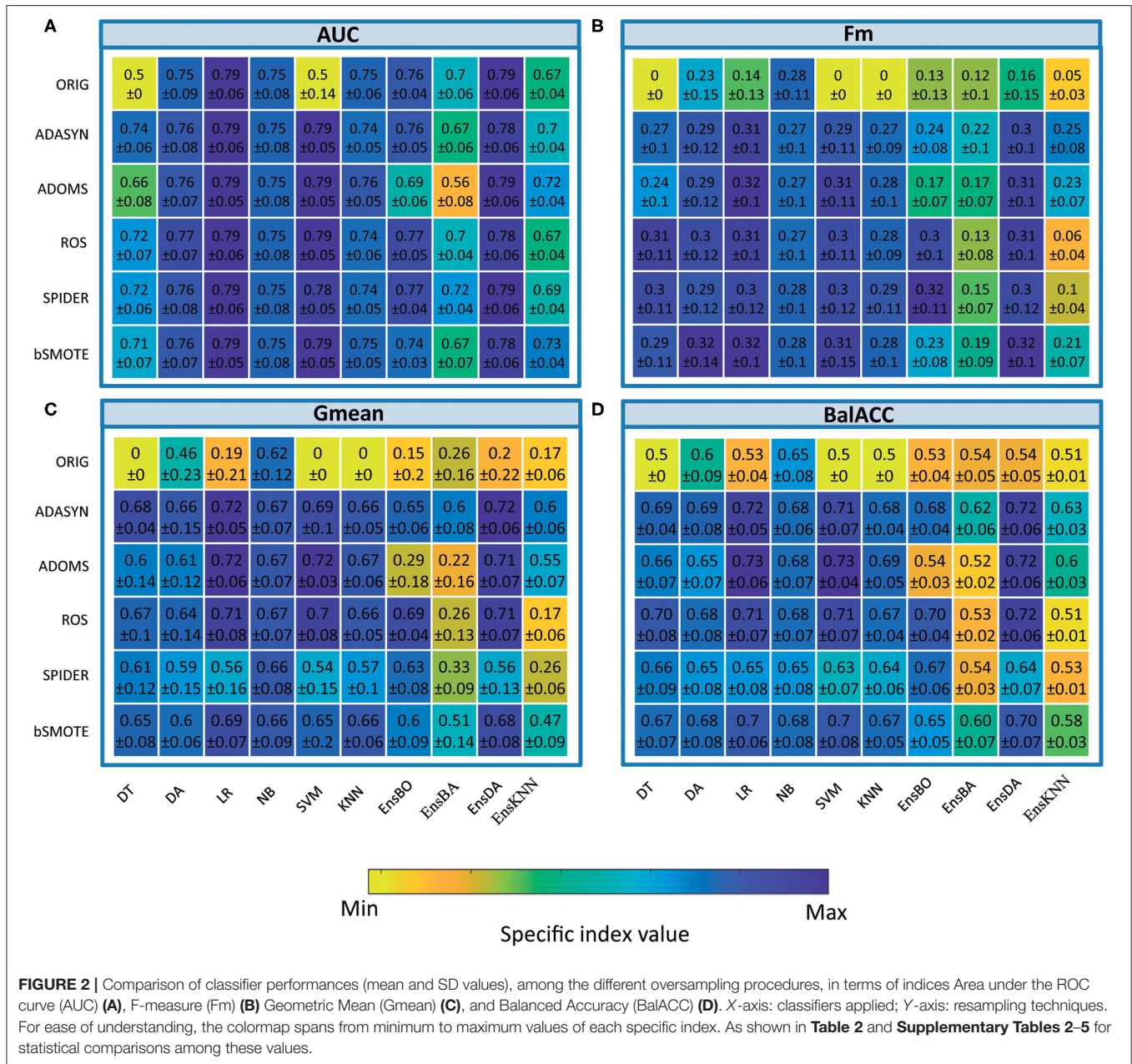
$$BalACC = \frac{TPr + TNr}{2}$$

Where TPr is the *true positive rate* (or *sensitivity*), TNr is the *true negative rate* (or *specificity*), and PPV is the *positive predicted value*, respectively, defined as:

$$TPr = \frac{TP}{TP + FN}$$

$$TNr = \frac{TN}{TN + FP}$$

$$PPV = \frac{TP}{TP + FP}$$



Note that TP, TN, FP, and FN stay for true positives, true negatives, false positives, and false negatives, respectively. For Fm we used $\beta = 1$, to assign equal importance to both TP and PPV.

All the analyses were performed using the *KEEL* software (Alcalá-Fdez et al., 2011) and the *Matlab classification learner* toolbox.

Statistical Analysis

To compare the different resampling techniques, Friedman’s test was applied to the four performances metrics AUC, Fm, Gmean, and BalACC (Friedman, 1937). When a significant difference among the group was found, Shaffer’s *post-hoc* test was applied for multiple comparisons (Shaffer, 1986). The alpha level for

statistical significance was set at 0.05, and the final adjusted *p*-values are used for the results. All the statistical comparisons were performed using SPSS (IBM Corp. Version 26.0. Armonk, NY, USA) and *KEEL* software.

Data are available from the corresponding authors upon request.

RESULTS

Oversampling

The average predicted performances in terms of AUC, Fm, Gmean, and BalACC are shown in **Figure 2**. For all 10 classifiers, the statistical results of the Friedman’s Test and related

TABLE 3 | Friedman's and *post-hoc* Shaffer's test for the *oversampling* techniques applied to the four performance measures: Area under the ROC Curve (AUC), F-measure (Fm), Geometric Mean (Gmean), and Balanced Accuracy (BalACC).

Oversampling		DT	DA	LR	NB	SVM	KNN	EnsBO	EnsBA	EnsDA	EnsKNN
Original vs. ADASYN	AUC	-				-		-			-
	Fm	-						-	-		-
	Gmean	-	-	-		-	-	-	-	-	-
	BalACC	-	-	-	-	-	-	-	-	-	-
Original vs. ADOMS	AUC					-			+		-
	Fm			-		-	-				
	Gmean	-		-		-	-			-	-
	BalACC	-		-	-	-	-			-	-
Original vs. ROS	AUC	-				-					
	Fm	-		-		-	-	-		-	-
	Gmean	-	-	-		-	-	-		-	
	BalACC	-	-	-	-	-	-	-		-	
Original vs. SPIDER	AUC	-									
	Fm	-		-		-	-	-		-	-
	Gmean							-			
	BalACC				-			-			
Original vs. bSMOTE	AUC					-					-
	Fm	-		-		-	-			-	-
	Gmean	-	-	-		-	-	-	-	-	-
	BalACC	-	-	-		-	-	-		-	-
ADASYN vs. ADOMS	AUC							+			
	Fm										
	Gmean							+	+		
	BalACC							+	+		
ADASYN vs. ROS	AUC										
	Fm								+		
	Gmean								+		+
	BalACC								+		+
ADASYN vs. SPIDER	AUC										
	Fm										
	Gmean			+						+	+
	BalACC			+						+	+
ADASYN vs. bSMOTE	AUC										
	Fm										
	Gmean										
	BalACC										
ADOMS vs. ROS	AUC							-	-		+
	Fm							-			
	Gmean							-			+
	BalACC							-			+
ADOMS vs. SPIDER	AUC							-	-		
	Fm							-			
	Gmean			+			+			+	
	BalACC			+			+	-		+	
ADOMS vs. bSMOTE	AUC										
	Fm										
	Gmean								-		
	BalACC								-		

(Continued)

TABLE 3 | Continued

Oversampling		DT	DA	LR	NB	SVM	KNN	EnsBO	EnsBA	EnsDA	EnsKNN
ROS vs. SPIDER	AUC										
	Fm										
	Gmean			+						+	
	BalACC			+						+	
ROS vs. bSMOTE	AUC										-
	Fm										
	Gmean								-		-
	BalACC										-
SPIDER vs. bSMOTE	AUC										
	Fm										
	Gmean										
	BalACC										

The 10 columns refer to the 10 classifiers models. The comparisons showing significant results are indicated with a “-” sign when the first algorithm (of the two compared in each row) was lower or with a “+” sign when it was higher than the second one. The rows without significant differences are not reported. Complete results with the p-values can be found in **Supplementary Tables 3-6**.

Shaffer’s *post-hoc* comparisons for AUC (a), Fm (b), Gmean (c), and BalACC (d) are shown in **Table 3**. Shaffer’s *post-hoc* comparisons have been indicated only when Friedman’s test resulted significantly. The sign “-” (respectively, “+”) indicates that the first algorithm has a lower (higher) value than the second one.

- *The area under the ROC curve*: Friedman’s test revealed significant differences among the pre-processing techniques only in five of the classifiers tested (DT, SVM, Ens_BO, Ens_BA, and Ens_KNN). For the two standard classifiers (DT and SVM), the *post-hoc* comparisons revealed differences only with respect to the original datasets, while no differences were present among the five oversampling techniques. Interestingly, for three of the four classical ensemble classifiers, none of the resampling techniques performed better than the original dataset. On the contrary, the ADOMS approach showed significantly lower AUC values than the other methods in both boosted and bagged ensemble classifiers. In the KNN ensemble, both original and ROS datasets reported the lowest performances (as shown in **Table 3** and **Supplementary Table 2**).
- *F-measure*: the significant differences have been revealed in 8 out of the 10 classifiers (DT, LR, SVM, KNN, EnsBO, EnsBA, EnsDA, and EnsKNN). The *post-hoc* comparisons showed the lower performance of the original dataset with respect to all resampling procedures in the six standard classifiers. In the ensemble both original and ADOMS had significantly lower Fm values than the other algorithms (as shown in **Table 3** and **Supplementary Table 3**).
- *Geometric Mean*: this metric exhibited more differences among the considered resampling approaches. All the classifiers except LR showed significant differences among the rebalancing approaches. In the standard classifiers and the EnsDA, the algorithm ADASYN, ADOMS, ROS, and bSMOTE performed better than both the original and SPIDER dataset.

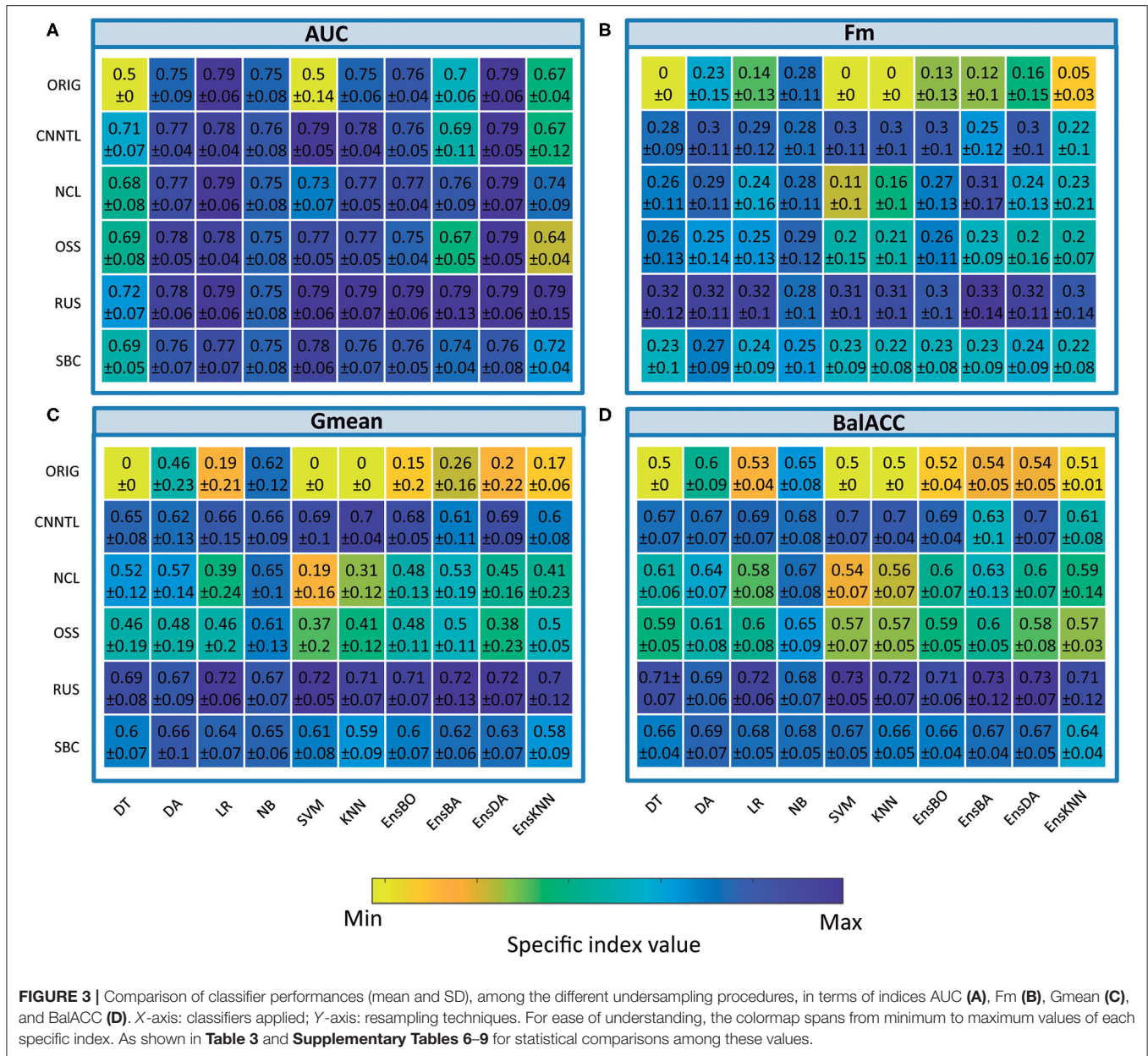
As for Fm, in Boosted and Bagged and KNN Ensemble ADOMS algorithm reported the lowest performance (as shown in **Table 3** and **Supplementary Table 4**).

- *Balanced Accuracy*: significant differences among the different resampling algorithms emerged for all the 10 classifiers. According to Shaffer’s *post-hoc* analysis, ADASYN, ADOMS, ROS, and bSMOTE reported better performances than the original and SPIDER datasets in the standard classifiers. In the EnsBO and EnsBA, no differences were found between the original and ADOMS data set, which performed worse than the other resampling procedures. In the EnsDA classifier, the resampling algorithms ADASYN, ADOMS, ROS, and bSMOTE showed higher BalACC than the original and SPIDER dataset. In EnsKNN classifier, showed similar results than EnsDA, except for ROS, which reported BalACC comparable with original and SPIDER (as shown in **Table 3** and **Supplementary Table 5**).

Undersampling

The average predicted performances of undersampling procedures in terms of AUC, Fm, Gmean, and BalACC are shown in **Figure 3**. For all the 10 classifiers, the statistical results of the Friedman’s Test and related Shaffer’s *post-hoc* comparisons for AUC (a), Fm (b), Gmean, (c), and BalACC (d) are shown in **Table 4**, respectively. Shaffer’s *post-hoc* comparisons have been indicated only when Friedman’s Test resulted significantly; The sign “-” (respectively “+”) indicates that the first algorithm has a lower (higher) value than the second one.

- *The area under the ROC curve*: significant differences among the pre-processing techniques are found in five of the classifiers tested (DT, SVM, KNN, EnsBA, and EnsKNN). In the DT classifier, all undersampling algorithms performed equally and better than the original one; in SVM, RUS,



and CNNTL performed better than the others, and in KNN only RUS showed improved AUC performances with respect to the original and all the other resampling techniques. In EnsBA and EnsKNN, significantly improved performances were achieved by NCL, RUS, and SBC (as shown in **Table 4** and **Supplementary Table 6**).

- *F-measure*: Friedman’s test revealed significant differences in 9 out of the 10 classifiers (all except NB). For standard classifiers, *post-hoc* comparisons showed the lower performance of the original dataset with respect to all resampling procedures except for SBC in DT classifier, NCL in SVM and KNN, and NCL, OSS, and SBC in LR classifier. As well as in standard classifiers, also in all the ensembles, the best performances were

achieved by RUS, followed by the CNNTL algorithm (as shown in **Table 4** and **Supplementary Table 7**).

- *Geometric Mean* showed significant differences among the considered approaches for all the classifiers, proving to be more suited than AUC and Fm in capturing the differences among the resampling approaches. RUS, SBC, and CNNTL showed the highest performances, with significantly higher Gmean than the original dataset in all the classifiers except NB. Moreover, RUS indicated significantly higher performances than NCL and OSS (**Table 4** and **Supplementary Table 8**).
- *Balanced Accuracy* showed very similar patterns with respect to Gmean, denoting differences for all the classifiers. According to Shaffer’s *post-hoc* analysis, CNNTL, RUS,

and SBC perform significantly better than the original dataset and the NCL and OSS resampling approaches, being RUS the best algorithms (as shown in **Table 4** and **Supplementary Table 9**).

Ensemble Methods for Imbalanced Domain

To compare the ensemble methods, we considered the two indices Gmean and BalACC since they have been shown to better capture the differences among the algorithms, as reported in the previous section.

Figure 4 shows the average ranking value for each of the proposed ensemble approaches, for both Gmean and BalACC. Corresponding results according to *post-hoc* Shaffer's test, comparing the seven approaches (original dataset and six ensembles) can be found in **Table 5**. According to Friedman's test, both the measures indicated significant differences among these techniques (Gmean: $p < 0.00001$; BalACC: $p < 0.00001$). A *post-hoc* analysis pointed out that DATABoost and SMOTEBag did not improve the performances with respect to the original dataset, and that SMOTEBag, OVERBag showed higher BalACC than the original data but no differences in terms of Gmean. On the contrary, RUSBoost and UNDERBag showed significantly better performances than all the other algorithms, being UNDERBag the best one (**Table 5**).

Since in the previous section we used classical ensemble classifiers combined with a rebalancing pre-processing step, we also compared the one with better performances (EnsDA, after ADASYN and RUS resampling) with the best algorithm of the modified ensemble family UNDERBag. Interestingly, EnsDA, with both ADASYN and RUS pre-processing, showed significantly higher Gmean and BalACC than the UNDER_Ba approach ($p < 0.00519$ for ADASYN+Ens_DA vs. UNDERBag, and $p < 0.00104$ for RUS+Ens_DA vs. UNDERBag, for both Gmean and BalACC). **Figure 5** represents the comparison among these three methods, expressed in terms of ranking values.

Sensitivity and Specificity

To clarify the effective use of the proposed approach to EZ identification, we reported sensitivity and specificity for the different techniques tested in the study. Since ensemble approaches showed significantly lower performances than resampling in terms of performances metrics (as indicated in the previous paragraph), only the sensitivity and specificity of the latter were further analyzed. **Figure 6** shows the boxplots indicating the values of sensitivity (full-color boxes) and specificity (horizontal lines boxes) for the original dataset compared with the five oversampling (**Figure 6A**) and the five undersampling approaches (**Figure 6B**). Each box represents the variability among the 10 classification models. All sensitivity and specificity values are reported in **Table 6**. Such results confirmed the main evidence obtained by the other performance metrics: (i) original data were not able to provide a good classification, since all the models tended to classify the whole set of leads as non-EZ (sensitivity ≈ 0 ; specificity ≈ 1), confirming the biased classification toward the majority non-EZ class; (ii) oversampling improved classification performances,

especially in terms of sensitivity. The Adasyn method provided the highest combination of both values (sensitivity and specificity >0.7) and the lowest variability of performances among the classification models. The ADOMS method showed average performances comparable with ADASYN, but much more variability with respect to the model choice. The SPIDER method was the least effective approach to improve the performances; (iii) Some undersampling approaches improved the classification performances, but with a strong variability among the different methods. NCL and OSS show results comparable to the original dataset. The RUS method provided the highest values of both sensitivity and specificity, comparable with the ADASYN approach. Interestingly, the SBC showed the highest sensibility values (≈ 0.9), even if associated with a less balanced specificity.

Figure 7 shows the visualization of the surgical 3D scene for a representative patient (pt2), such as an indication of the resected zone (blue area), true EZ and non-EZ leads, and the EZ and non-EZ classification provided by the RUS + EnsDA method.

DISCUSSION

Machine learning approaches are being increasingly applied to the field of epilepsy, and specifically in the different datasets from neurophysiological recordings (Abbasi and Goldenholz, 2019). In this context, it is quite common to cope with the imbalanced datasets characterized by uneven distribution between majority and minority classes, which can lead to worse classification performances.

This is the case of the EZ localization in the pre-surgical planning to achieve seizure freedom after surgical resection of the EZ. One assessed clinical practice is the exploration through intracranial EEG recordings (SEEG) (Cardinale et al., 2019) combined with the visual analysis and advanced signal processing methods able to extract quantitative indexes to support the correct EZ localization (Bartolomei et al., 2017).

Intentionally, to sample a wide region of the epileptic brain, the explored brain regions are much wider than the true EZ, thus resulting in an imbalanced class distribution between EZ and non-EZ contacts, with the EZ being the most important class to be correctly identified to reduce or remove seizures, being the minority class. This led the classifier to be biased toward the majority (non-EZ) class.

Starting from the evidence that network analysis of interictal SEEG recordings could be very useful in support of the EZ localization (Varotto et al., 2012; Vlachos et al., 2017; Lagarde et al., 2018), in this study we demonstrated that the combination of supervised machine learning with appropriate data resampling approach can strongly improve its potential. For this reason, the idea of applying resampling techniques in the field of EZ localization should be taken into consideration.

At present, no study investigated the effect of imbalance domains on the performance of EZ localization methodologies. The previous studies demonstrated that the application of rebalancing techniques could strongly improve the classification of EEG signals for epilepsy diagnosis (Haldar et al., 2019; Kaur

TABLE 4 | Friedman's and *post-hoc* Shaffer's test for the *undersampling* techniques applied to the four performance measures: AUC, Fm, Gmean, and BalACC.

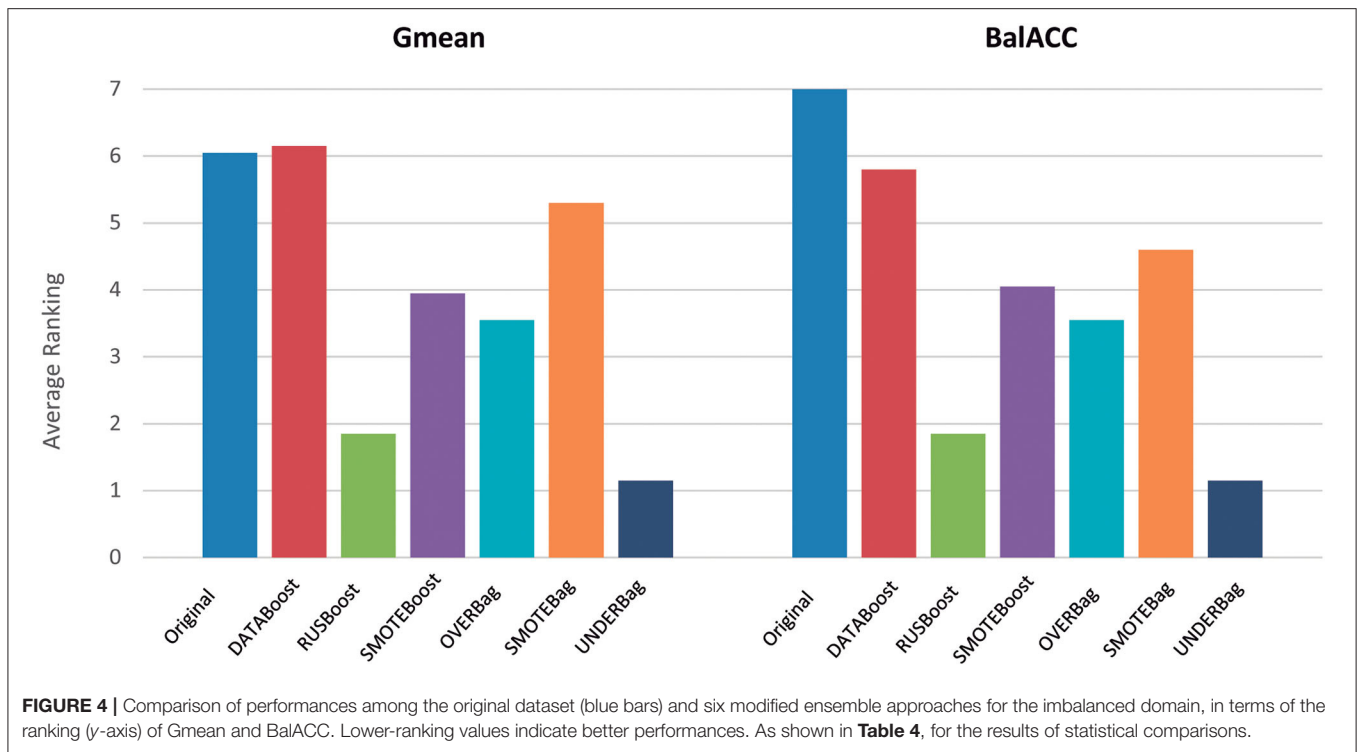
Undersampling		DT	DA	LR	NB	SVM	KNN	EnsBO	EnsBA	EnsDA	EnsKNN
Original vs. <i>CNNTL</i>	AUC	-				-					
	Fm	-		-		-	-	-		-	-
	Gmean	-	-	-		-	-	-	-	-	-
	BalACC	-	-	-	-	-	-	-		-	-
Original vs. <i>NCL</i>	AUC	-									
	Fm	-						-	-	-	
	Gmean										
	BalACC										
Original vs. <i>OSS</i>	AUC	-									
	Fm	-				-	-				-
	Gmean										
	BalACC										
Original vs. <i>RUS</i>	AUC	-				-	-				-
	Fm	-		-		-	-	-	-	-	-
	Gmean	-	-	-		-	-	-	-	-	-
	BalACC	-	-	-		-	-	-	-	-	-
Original vs. <i>SBC</i>	AUC	-				-					
	Fm					-	-				-
	Gmean	-		-		-	-	-	-	-	-
	BalACC	-		-		-	-	-	-	-	-
<i>CNNTL</i> vs. <i>NCL</i>	AUC					+			-		-
	Fm					+	+				-
	Gmean			+		+	+	+		+	
	BalACC			+		+	+				
<i>CNNTL</i> vs. <i>OSS</i>	AUC										
	Fm										
	Gmean				+	+	+	+		+	
	BalACC		+		+	+	+	+		+	
<i>CNNTL</i> vs. <i>RUS</i>	AUC								-		-
	Fm										
	Gmean										
	BalACC										
<i>CNNTL</i> vs. <i>SBC</i>	AUC								-		-
	Fm							+			
	Gmean										
	BalACC										
<i>NCL</i> vs. <i>OSS</i>	AUC								+		+
	Fm										
	Gmean										
	BalACC										
<i>NCL</i> vs. <i>RUS</i>	AUC					-					-
	Fm					-	-				-
	Gmean	-		-		-	-	-	-	-	-
	BalACC	-		-		-	-	-	-	-	-
<i>NCL</i> vs. <i>SBC</i>	AUC										
	Fm										
	Gmean										
	BalACC			-							

(Continued)

TABLE 4 | Continued

Undersampling		DT	DA	LR	NB	SVM	KNN	EnsBO	EnsBA	EnsDA	EnsKNN
OSS vs. RUS	AUC								-		-
	Fm										
	Gmean	-	-	-		-	-	-	-	-	-
	BalACC	-	-	-		-	-	-	-	-	-
OSS vs. SBC	AUC								-		-
	Fm										
	Gmean										
	BalACC										
RUS vs. SBC	AUC										
	Fm				+						
	Gmean										
	BalACC										

The 10 columns refer to the 10 classifiers models. The comparisons showing significant results are indicated with a “-” sign when the first algorithm (of the two compared in each row) was lower or with a “+” sign when it was higher than the second one. The rows without significant differences are not reported. Complete results with the p-values can be found in **Supplementary Tables 3-6**.



et al., 2020) and automatic seizure detection (Cosgun et al., 2019; Romaiisa et al., 2019; Masum et al., 2020). However, in most of them, the well-known and assessed resampling techniques belonging to the SMOTE family were applied, and systematic comparison with other possible approaches was missing.

In this study, we compared five oversampling and five undersampling procedures and tested the resulting rebalanced datasets with 10 different machine learning classifiers. Moreover, we also tested six specific ensemble methods properly modified

for imbalanced domain and belonging to data variation-based ensemble.

Our study focuses on identifying the best resampling and classification approach to support the classification of brain regions as EZ or non-EZ, using the indexes derived from connectivity and graph-theory analysis of interictal SEEG recording as features. The selection of the nine graph-theory-based indexes used as input features of the classifiers was based on the preliminary analysis we performed, showing that the combination of these indexes was the most appropriate

TABLE 5 | Shaffer's test for the ensemble approaches for the imbalance domain.

Ensemble		AUC		Fm		Gmean		BalACC
Original vs. <i>DATABoost</i>	–	0.000	–	0.014		1.895		1.285
Original vs. <i>RUSBoost</i>		2.344		0.789	–	0.000	–	0.000
Original vs. <i>SMOTEBoost</i>		0.165		0.423		0.297	–	0.034
Original vs. <i>OVERBag</i>		0.555		1.285		0.106	–	0.005
Original vs. <i>SMOTEBag</i>	–	0.040	–	0.006		1.895		0.143
Original vs. <i>UNDERBag</i>		2.344	–	0.000	–	0.000	–	0.000
<i>DATABoost</i> vs. <i>RUSBoost</i>	+	0.000		0.789	–	0.000	–	0.001
<i>DATABoost</i> vs. <i>SMOTEBoost</i>		0.555		1.285		0.251		0.491
<i>DATABoost</i> vs. <i>OVERBag</i>		0.218		0.372		0.078		0.199
<i>DATABoost</i> vs. <i>SMOTEBag</i>		1.499		1.814		1.895		1.285
<i>DATABoost</i> vs. <i>UNDERBag</i>	+	0.000		0.298	–	0.000	–	0.000
<i>RUSBoost</i> vs. <i>SMOTEBoost</i>		0.165		1.814		0.297		0.205
<i>RUSBoost</i> vs. <i>OVERBag</i>		0.555		1.814		0.549		0.549
<i>RUSBoost</i> vs. <i>SMOTEBag</i>	+	0.040		0.701	+	0.005	+	0.049
<i>RUSBoost</i> vs. <i>UNDERBag</i>		2.344	–	0.002		1.895		1.406
<i>SMOTEBoost</i> vs. <i>OVERBag</i>		2.344		1.516		1.895		1.406
<i>SMOTEBoost</i> vs. <i>SMOTEBag</i>		2.344		1.031		0.974		1.406
<i>SMOTEBoost</i> vs. <i>UNDERBag</i>	–	0.024	–	0.006		0.056	–	0.034
<i>OVERBag</i> vs. <i>SMOTEBag</i>		1.663		0.298		0.491		1.285
<i>OVERBag</i> vs. <i>UNDERBag</i>		0.091	–	0.000		0.143		0.143
<i>SMOTEBag</i> vs. <i>UNDERBag</i>	–	0.003		0.423	–	0.000	–	0.005

Red color indicates the *p*-values with significant differences according to Shaffer's post-hoc ($p < 0.05$); the sign "–" (respectively "+") indicates that the first algorithm has a lower (higher) value than the second one.

to achieve the best EZ classification. In the contest of EZ localization, despite the early application of several other signal processing approaches for feature extraction, such as working in the frequency domain or by non-linear analysis, network analysis started only recently to be employed based on the evidence that focal epilepsy is a network disease. However, most of these recent network studies normally focus only on the connectivity analysis that is rarely combined with the pre-processing approaches, due to the huge amount of data to be processed. For this reason, in this study, we mainly focused on presenting pre-processing, in combination with a few of such feature extraction and connectivity measures in the literature, to provide evidence of and support for a proper pre-processing method in this context.

Regarding oversampling, all five approaches reported improved performances with respect to the original dataset. The differences among the five oversampling approaches varied according to the considered classifiers.

Adaptive Synthetic Sampling resulted to be the most robust approach among the classifiers. ADOMS was the less robust and most sensitive to the choice of classifier, being comparable or even slightly better than ADASYN in LR, SVM, KNN, EnsDA, and EnsKNN, while as bad as the original dataset in DA, EnsBO, and EnsBA. SPIDER was the least effective, with performances significantly worse than the other approaches and comparable with the original dataset for some classifiers, especially the classical ensemble family.

Regarding undersampling, all the approaches appeared to be less influenced by the classifier choice than the oversampling.

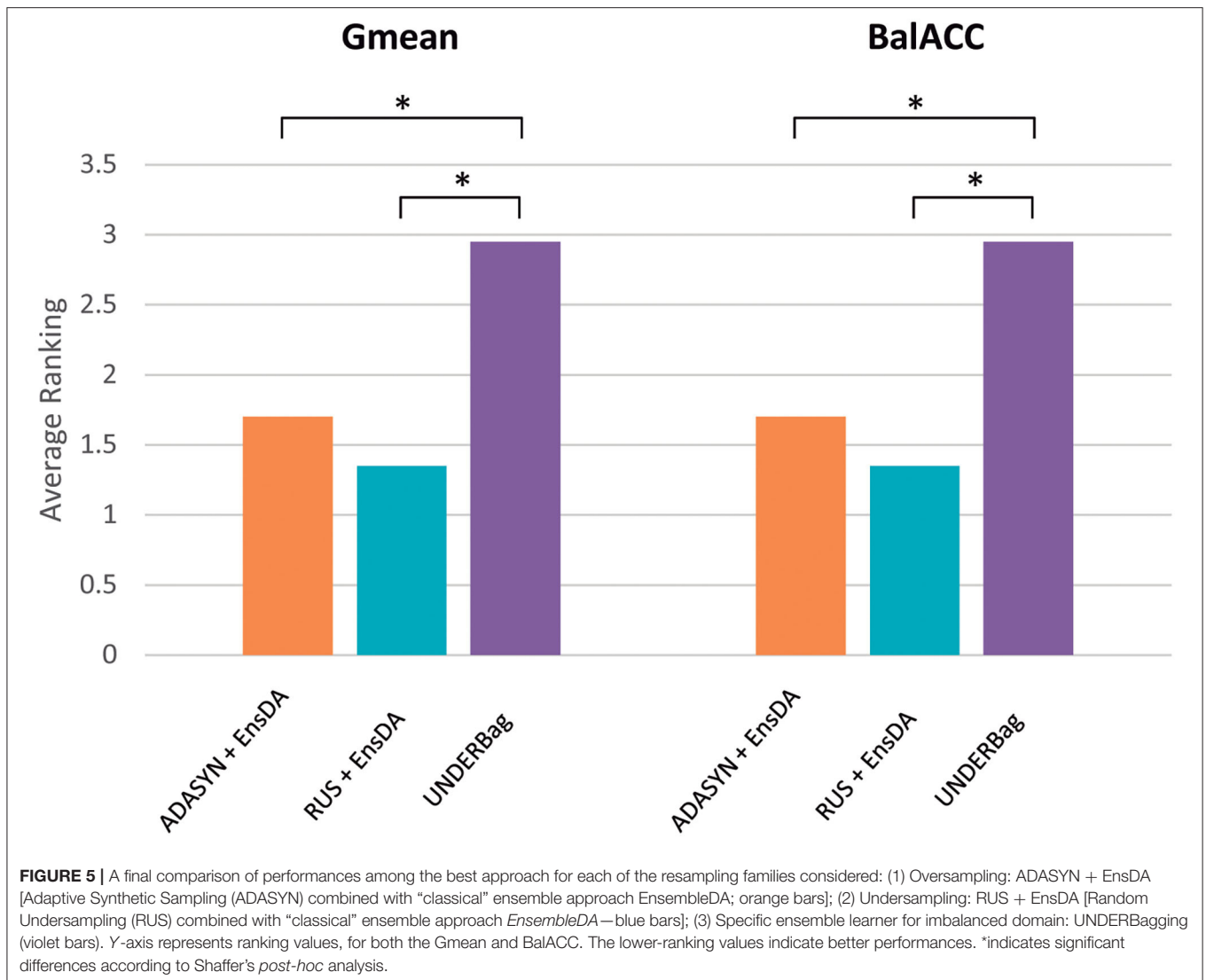
Two of the proposed methods, NCL and OSS, did not improve the classification performances with respect to the original data. The other approaches were significantly better than original data, with RUS, the simplest of the proposed methods, being the best one.

Interestingly RUS showed higher, even not significant, performances than the best oversampling approach, ADASYN.

The resampling technique is not the only family to cope with the imbalanced domain. A wide number of approaches exist to deal with this problem, which can be mainly categorized as data-level or algorithmic-level approaches (López et al., 2013). Rebalancing belongs to the data-level approaches, in which data are pre-processed before the classification (Lee, 2014). On the contrary, in the algorithmic-level ones, the classification algorithm is modified to deal with the imbalanced nature (Barandela et al., 2003a). The cost-sensitive approaches combine both the data and algorithmic levels, by assigning different misclassification costs for the two classes and modify the classification algorithm to minimize the higher misclassification cost (Domingos, 1999; Zhou and Liu, 2006; Sun et al., 2007).

The main limitation of cost-sensitive approaches is the need of defining the correct misclassification costs for the two classes, which may not be so clear in many clinical problems, as in our case.

In this paper, we focused on the rebalancing techniques since they can be quite easily implemented, and are independent of the underlying classifiers, which can be an advantage in problems where the selection of the most appropriate classifier is not clear (Batista et al., 2004; Batuwita and Palade, 2010).



In addition, several modifications of ensemble methods for the imbalanced domain have been proposed (Rokach, 2010), both working at data-level approach, through the data pre-processing before each step of the ensemble classification (Breiman, 1996; Freund and Schapire, 1997; Kuncheva, 2014), or with algorithmic-level cost-sensitive modification (Sun et al., 2007).

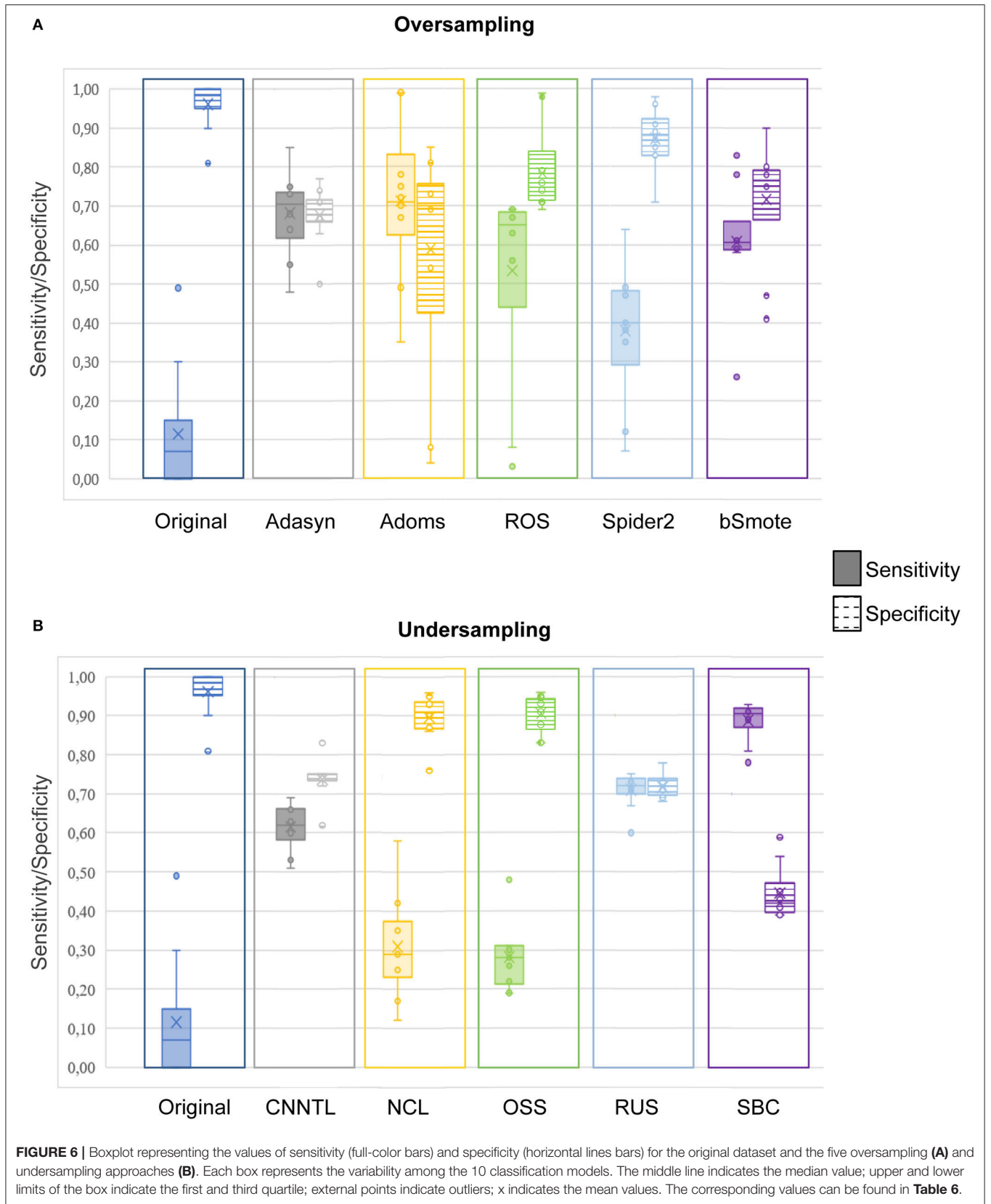
As part of the data-level approaches, we considered and tested, in this study, six different data-level ensemble algorithms. As reported in a previous study (Galar et al., 2012), we found that the simplest algorithms, UNDERBag and RUSBoost emerged as the best ensemble methods, while offering lower computation costs.

Interestingly, when compared these results with those obtained by a standard single-step resampling approach combined with a classical ensemble algorithm, we found significantly higher performances in the latter family, in particular for the combination (ADASYN + EnsDA and RUS + EnsDA).

This highlights again that the simplest algorithms guarantee high performances, and that their very low computational complexity can be a strong advantage toward routine clinical applications.

It is important to notice that the performances of the different resampling techniques are strongly influenced by the choice of the classifier. This highlights that the selection of the resampling approach for a specific dataset should always take into consideration the choice of the classifier.

Regarding the measure to assess and compare the performances, in this study we applied four measures considered most appropriate to deal with imbalanced classification: AUC, Fm, Gmean, and BalACC (Bekkar et al., 2013). Several studies already highlighted that the choice of the proper evaluation measures for model assessment is one of the most complex issues faced in the imbalanced data learning context and how the application of more standard measures, such as accuracy, could lead to erroneous interpretations and biased classification (Weiss, 2004).



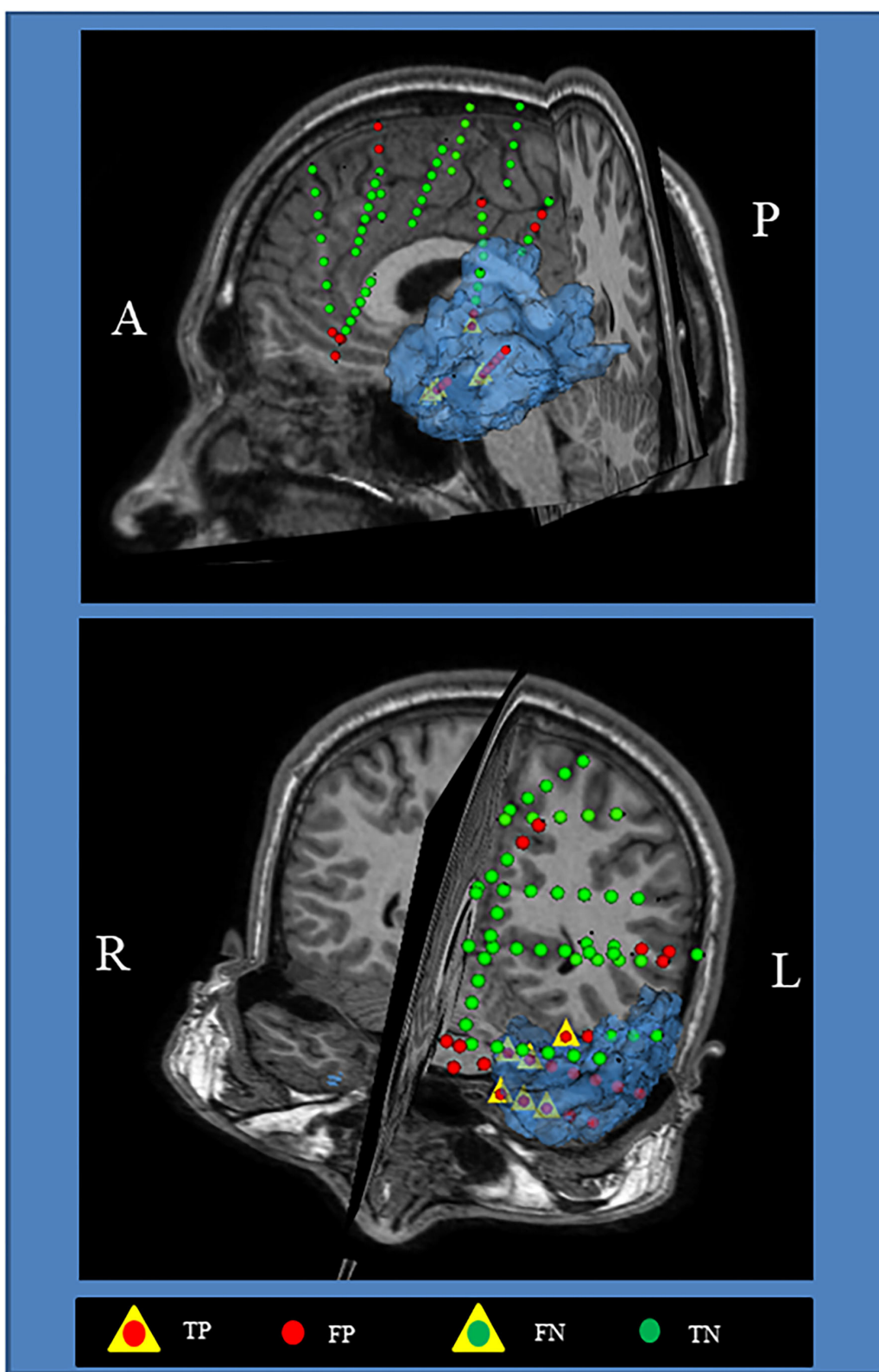


FIGURE 7 | Visualization of the surgical 3D scene for a representative patient (pt2). The Blue area indicates the final resected zone. Red and green dot points indicate leads classified as epileptogenic zone (EZ) A, anterior; L, left; P, posterior; R, right.

TABLE 6 | The Sensitivity (Sens) and Specificity (Spec) values for oversampling and undersampling techniques.

	Oversampling											
	Orig		Adasyn		Adoms		ROS		Spider2		bSmote	
	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
DT	0.00	1.00	0.75	0.63	0.78	0.54	0.63	0.76	0.49	0.83	0.59	0.75
DA	0.30	0.90	0.64	0.74	0.49	0.81	0.56	0.79	0.47	0.83	0.58	0.78
LR	0.08	0.98	0.73	0.71	0.72	0.74	0.68	0.74	0.40	0.89	0.61	0.79
NB	0.49	0.81	0.68	0.68	0.67	0.69	0.67	0.69	0.64	0.71	0.62	0.73
SVM	0.00	1.00	0.73	0.68	0.75	0.70	0.69	0.72	0.35	0.91	0.60	0.80
KNN	0.00	1.00	0.68	0.67	0.68	0.70	0.63	0.71	0.40	0.87	0.59	0.75
EnsBO	0.06	0.99	0.85	0.50	0.99	0.08	0.67	0.72	0.48	0.85	0.83	0.47
EnsBA	0.10	0.97	0.55	0.68	0.99	0.04	0.08	0.98	0.12	0.96	0.78	0.41
EnsDA	0.09	0.98	0.73	0.71	0.70	0.73	0.69	0.74	0.38	0.90	0.61	0.78
EnsKNN	0.03	0.99	0.48	0.77	0.35	0.85	0.03	0.99	0.07	0.98	0.26	0.90
Mean	0.12	0.96	0.68	0.68	0.71	0.59	0.53	0.78	0.38	0.87	0.61	0.72
St. Dev	0.16	0.06	0.11	0.07	0.20	0.29	0.26	0.11	0.17	0.08	0.15	0.15

	Undersampling											
	Orig		CNNTL		NCL		OSS		RUS		SBC	
	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
DT	0.00	1.00	0.61	0.74	0.35	0.87	0.28	0.91	0.71	0.71	0.92	0.39
DA	0.30	0.90	0.51	0.83	0.42	0.86	0.30	0.91	0.60	0.78	0.78	0.59
LR	0.08	0.98	0.63	0.75	0.26	0.91	0.28	0.92	0.71	0.74	0.91	0.45
NB	0.49	0.81	0.61	0.74	0.58	0.76	0.48	0.83	0.67	0.69	0.81	0.54
SVM	0.00	1.00	0.69	0.72	0.12	0.96	0.19	0.96	0.75	0.70	0.92	0.41
KNN	0.00	1.00	0.67	0.74	0.17	0.95	0.19	0.95	0.74	0.70	0.93	0.39
EnsBO	0.06	0.99	0.64	0.75	0.29	0.91	0.26	0.93	0.73	0.70	0.92	0.41
EnsBA	0.10	0.97	0.53	0.74	0.36	0.90	0.31	0.88	0.74	0.73	0.90	0.43
EnsDA	0.09	0.98	0.66	0.74	0.29	0.91	0.22	0.94	0.71	0.74	0.90	0.45
EnsKNN	0.03	0.99	0.60	0.62	0.25	0.93	0.31	0.83	0.74	0.68	0.89	0.40
Mean	0.12	0.96	0.62	0.74	0.31	0.90	0.28	0.91	0.71	0.72	0.89	0.45
St. Dev	0.16	0.06	0.06	0.05	0.13	0.06	0.08	0.05	0.05	0.03	0.05	0.07

Single values for each of the 10 classifier models, as well as mean and standard deviation (St.Dev.) are indicated.

These four measures provided complementary results and to properly evaluate the performances of different approaches, it is important to take into account the combination of them, especially considering which aspect is more important in the specific problem we are facing. Particularly, in this case, we noticed that AUC and Fm did not completely capture differences in the model performances. On the other side, as already described in another paper (Luque et al., 2019), Gmean and BalACC appear to be good performance metrics when the main focus is to maximize sensitivity, without losing too much specificity.

DATA AVAILABILITY STATEMENT

Data are available from the corresponding authors upon request. Requests to access these datasets should be directed to giulia.varotto@istituto-besta.it.

ETHICS STATEMENT

The study was approved by the Ethics Committee of the Fondazione IRCCS Istituto Neurologico Carlo Besta of Milan and was carried out in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. All of the subjects gave their written informed consent before being included in the study.

AUTHOR CONTRIBUTIONS

GV: designed and conceptualized the study, analyzed and interpreted the data, and drafted the manuscript for intellectual content. GS: contributed to design the study, analyzed the data, and contributed to draft and revise the manuscript. LT and FG: major role in the acquisition of data and contributed to revise the manuscript. SF and FP: interpreted the data and contributed to

draft and revise the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the DESIRE (Strategies for Innovative Research to Improve Diagnosis, Prevention, and Treatment in children with difficult to treat epilepsy), an FP7 funded project (Grant Agreement No. 602531), from the European Commission, and the Grants Nos. RF-2011-02350578 and RF-2010-2319316 from the Italian Ministry of Health.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fninf.2021.715421/full#supplementary-material>

Supplementary Figure 1 | A subset of stereo-electroencephalography (SEEG) traces recorded from pt2, and corresponding adjacency matrices for the first 3 of the 36 epochs analyzed.

Supplementary Table 1 | Description of the set of graph-theory based centrality measures used in this study.

Supplementary Table 2 | Friedman and *post-hoc* Shaffer test for the *oversampling* techniques with *AUC* measure. Shaffer *post-hoc* comparisons have been indicated only when Friedman test resulted significant (p -values in the first line). Red color indicates p -values with significant differences according to shaffer *post-hoc* ($p < 0.05$); “–” (respectively “+”) indicates that the first algorithm has lower (higher) value than the second one.

Supplementary Table 3 | Friedman and *post-hoc* Shaffer test for the *oversampling* techniques with *Fm* measure. Shaffer *post-hoc* comparisons have been indicated only when Friedman test resulted significant (p -values in the first line). Red color indicates p -values with significant differences according to shaffer *post-hoc* ($p < 0.05$); “–” (respectively “+”) indicates that the first algorithm has lower (higher) value than the second one.

REFERENCES

- Abbasi, B., and Goldenholz, D. M. (2019). Machine learning applications in epilepsy. *Epilepsia* 60, 2037–2047. doi: 10.1111/epi.16333
- Acharya, U. R., Hagiwara, Y., and Adeli, H. (2018). Automated seizure prediction. *Epilepsy Behav.* 88, 251–261 doi: 10.1016/j.yebeh.2018.09.030
- Adkinson, J. A., Karumuri, B., Hutson, T. N., Liu, R., Alamoudi, O., Vlachos, I., et al. (2019). Connectivity and centrality characteristics of the epileptogenic focus using directed network analysis. *IEEE Trans. Neural Syst. Rehabil. Eng.* 27, 22–30. doi: 10.1109/TNSRE.2018.2886211
- Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., et al. (2011). KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *J. Mult. Log. Soft Comput.* 17, 255–287. Available online at: <http://www.oldcitypublishing.com/journals/mvlsc-home/mvlsc-issue-contents/mvlsc-volume-17-number-2-3-2011/mvlsc-17-2-3-p-255-287/>
- Ali, A., Shamsuddin, S. M., and Ralescu, A. L. (2015). Classification with class imbalance problem: a review. *Int. J. Adv. Soft Comput. Appl.* 7, 176–204.
- Armañanzas, R., Alonso-Nanclares, L., DeFelipe-Oroquieta, J., Kastanauskaitė, A., de Sola, R. G., DeFelipe, J., et al. (2013). Machine learning approach for the outcome prediction of temporal lobe epilepsy surgery. *PLoS ONE*. 8:e62819. doi: 10.1371/journal.pone.0062819
- Azami, M., El Hammers, A., Jung, J., Costes, N., Bouet, R., and Lartizien, C. (2016). Detection of lesions underlying intractable epilepsy on t1-weighted mri as an outlier detection problem. *PLoS ONE* 11:e0161498. doi: 10.1371/journal.pone.0161498

Supplementary Table 4 | Friedman and *post-hoc* Shaffer test for the *oversampling* techniques with *Gmean* measure. Shaffer *post-hoc* comparisons have been indicated only when Friedman test resulted significant (p -values in the first line). Red color indicates p -values with significant differences according to shaffer *post-hoc* ($p < 0.05$); “–” (respectively “+”) indicates that the first algorithm has lower (higher) value than the second one.

Supplementary Table 5 | Friedman and *post-hoc* Shaffer test for the *oversampling* techniques with *BalACC* measure. Shaffer *post-hoc* comparisons have been indicated only when Friedman test resulted significant (p -values in the first line). Red color indicates p -values with significant differences according to shaffer *post-hoc* ($p < 0.05$); “–” (respectively “+”) indicates that the first algorithm has lower (higher) value than the second one.

Supplementary Table 6 | Friedman and *post-hoc* Shaffer test for the *oversampling* techniques with *AUC* measure. Shaffer *post-hoc* comparisons have been indicated only when Friedman test resulted significant (p -values in the first line). Red color indicates p -values with significant differences according to shaffer *post-hoc* ($p < 0.05$); “–” (respectively “+”) indicates that the first algorithm has lower (higher) value than the second one.

Supplementary Table 7 | Friedman and *post-hoc* Shaffer test for the *undersampling* techniques with *Fm* measure. Shaffer *post-hoc* comparisons have been indicated only when Friedman test resulted significant (p -values in the first line). Red color indicates p -values with significant differences according to shaffer *post-hoc* ($p < 0.05$); “–” (respectively “+”) indicates that the first algorithm has lower (higher) value than the second one.

Supplementary Table 8 | Friedman and *post-hoc* Shaffer test for the *undersampling* techniques with *Gmean* measure. Shaffer *post-hoc* comparisons have been indicated only when Friedman test resulted significant (p -values in the first line). Red color indicates p -values with significant differences according to shaffer *post-hoc* ($p < 0.05$); “–” (respectively “+”) indicates that the first algorithm has lower (higher) value than the second one.

Supplementary Table 9 | Friedman and *post-hoc* Shaffer test for the *undersampling* techniques with *BalACC* measure. Shaffer *post-hoc* comparisons have been indicated only when Friedman test resulted significant (p -values in the first line). Red color indicates p -values with significant differences according to shaffer *post-hoc* ($p < 0.05$); “–” (respectively “+”) indicates that the first algorithm has lower (higher) value than the second one.

- Barandela, R., Sánchez, J. S., García, V., and Rangel, E. (2003a). Strategies for learning in class imbalance problems. *Pattern Recognit.* 36, 849–851. doi: 10.1016/S0031-3203(02)00257-1
- Barandela, R., Sánchez, J. S., and Valdovinos, R. M. (2003b). New Applications of ensembles of classifiers. *Pattern Anal. Appl.* 6, 245–256. doi: 10.1007/s10044-003-0192-z
- Bartolomei, F., Lagarde, S., Wendling, F., McGonigal, A., Jirsa, V., Guye, M., et al. (2017). Defining epileptogenic networks: contribution of SEEG and signal analysis. *Epilepsia* 58, 1131–1147. doi: 10.1111/epi.13791
- Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* 6, 20–29. doi: 10.1145/1007730.1007735
- Batuwita, R., and Palade, V. (2010). “Efficient resampling methods for training support vector machines with imbalanced datasets,” in *Proceedings of the International Joint Conference on Neural Networks*, 1–8. doi: 10.1109/IJCNN.2010.5596787
- Bekkar, M., Djemaa, H. K., and Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *J. Inf. Eng. Appl.* 3, 2224–5782. Available online at: <https://www.iiste.org/Journals/index.php/JIEA/article/view/7633>
- Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24, 123–140. doi: 10.1007/BF00058655
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324
- Bulacio, J. C., Jehi, L., Wong, C., Gonzalez-Martinez, J., Kotagal, P., Nair, D., et al. (2012). Long-term seizure outcome after resective surgery in

- patients evaluated with intracranial electrodes. *Epilepsia* 53, 1722–1730. doi: 10.1111/j.1528-1167.2012.03633.x
- Cardinale, F., Rizzi, M., Vignati, E., Cossu, M., Castana, L., d'Orto, P., et al. (2019). Stereoelectroencephalography: retrospective analysis of 742 procedures in a single centre. *Brain* 142, 2688–2704. doi: 10.1093/brain/awz196
- Chawla, N. V. (2009). “Data mining for imbalanced datasets: an overview,” in *Data Mining and Knowledge Discovery Handbook*, eds O. Maimon, and L. Rokach (Boston, MA: Springer). doi: 10.1007/978-0-387-09823-4_45
- Chawla, N. V., Lazarevic, A., Hall, L. O., and Bowyer, K. W. (2003). “SMOTEBoost: improving prediction of the minority class in boosting,” in *Knowledge Discovery in Databases. Lecture Notes in Computer Science, Vol. 2838*, eds N. Lavrač, N. D. Gamberger, L. Todorovski, H. Blockeel (Berlin: Springer). doi: 10.1007/978-3-540-39804-2_12
- Cosgun, E., Celebi, A., and Gullu, M. K. (2019). “Epileptic seizure prediction for imbalanced datasets,” in *Medical Technologies Congress (TIPTEKNO)* (Izmir), 1–4. doi: 10.1109/TIPTEKNO.2019.8895137
- Daoud, H., and Bayoumi, M. A. (2019). Efficient epileptic seizure prediction based on deep learning. *IEEE Trans. Biomed. Circuits Syst.* 13, 804–813. doi: 10.1109/TBCAS.2019.2929053
- Dian, J. A., Colic, S., Chinvarun, Y., Carlen, P. L., and Bardakjian, B. L. (2015). Identification of brain regions of interest for epilepsy surgery planning using support vector machines. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2015, 6590–6593. doi: 10.1109/EMBC.2015.7319903
- Domingos, P. (1999). “MetaCost: a general method for making classifiers cost-sensitive,” in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data* (San Diego, CA), 155–164. doi: 10.1145/312129.312220
- Elahian, B., Yeasin, M., Mudigoudar, B., Wheless, J. W., and Babajani-Feremi, A. (2017). Identifying seizure onset zone from electrocorticographic recordings: a machine learning approach based on phase locking value. *Seizure* 51, 35–42. doi: 10.1016/j.seizure.2017.07.010
- Engel, J. (1993). Update on surgical treatment of the epilepsies: summary of the second international palm desert conference on the surgical treatment of the epilepsies (1992). *Neurology* 43, 1612–1617. doi: 10.1212/WNL.43.8.1612
- Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). “Foundations on imbalanced classification,” in *Learning From Imbalanced Data Sets* (Cham: Springer). doi: 10.1007/978-3-319-98074-4_2
- Fiest, K. M., Sauro, K. M., Wiebe, S., Patten, S. B., Kwon, C. S., Dykeman, J., et al. (2017). Prevalence and incidence of epilepsy. *Neurology* 88, 296–303. doi: 10.1212/WNL.0000000000003509
- Frank, B., Hurley, L., Scott, T. M., Olsen, P., Dugan, P., and Barr, W. B. (2018). Machine learning as a new paradigm for characterizing localization and lateralization of neuropsychological test data in temporal lobe epilepsy. *Epilepsy Behav.* 86, 58–65. doi: 10.1016/j.yebeh.2018.07.006
- Freund, Y., and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55, 119–139. doi: 10.1006/jcss.1997.1504
- Freund, Y., and Schapire, R. E. (1999). A short introduction to boosting. *J. Jpn. Soc. Artif. Intell.* 14, 771–780.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* 32:675. doi: 10.1080/01621459.1937.10503522
- Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. (2012). “A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches,” in *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol. 42* (Institute of Electrical and Electronics Engineers Inc.), 463–484. doi: 10.1109/TSMCC.2011.2161285
- Gleichgerricht, E., Munsell, B., Bhatia, S., Vandergrift, W. A., Rorden, C., McDonald, C., et al. (2018). Deep learning applied to whole-brain connectome to determine seizure control after epilepsy surgery. *Epilepsia* 59, 1643–1654. doi: 10.1111/epi.14528
- Goldenholz, D. M., Jow, A., Khan, O. I., Bagić, A., Sato, S., Auh, S., et al. (2016). Preoperative prediction of temporal lobe epilepsy surgery outcome. *Epilepsy Res.* 127, 331–338. doi: 10.1016/j.eplepsyres.2016.09.015
- Guo, H., and Viktor, H. L. (2004). Learning from imbalanced data sets with boosting and data generation. *ACM SIGKDD Explor. Newsl.* 6, 30–39. doi: 10.1145/1007730.1007736
- Haldar, S., Mukherjee, R., Chakraborty, P., Banerjee, S., Chaudhury, S., and Chatterjee, S. (2019). “Improved epilepsy detection method by addressing class imbalance problem,” in *IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 934–939. doi: 10.1109/IEMCON.2018.8614826
- Han, H., Wang, W. Y., and Mao, B. H. (2005). “Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning,” in *Advances in Intelligent Computing. Lecture Notes in Computer Science Vol. 3644*, eds D. S. Huang, X. P. Zhang, G. B. Huang (Berlin: Springer). doi: 10.1007/11538059_91
- He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). “ADASYN: adaptive synthetic sampling approach for imbalanced learning,” in *International Joint Conference on Neural Networks* (Nashville, TN: IEEE World Congress on Computational Intelligence), 1322–1328.
- He, H., and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* 21, 1263–1284. doi: 10.1109/TKDE.2008.239
- Jin, S. H., and Chung, C. K. (2017). Electrophysiological resting-state biomarker for diagnosing mesial temporal lobe epilepsy with hippocampal sclerosis. *Epilepsy Res.* 129, 138–145. doi: 10.1016/j.eplepsyres.2016.11.018
- Kassahun, Y., Perrone, R., De Momi, E., Berghöfer, E., Tassi, L., Canevini, M. P., et al. (2014). Automatic classification of epilepsy types using ontology-based and genetics-based machine learning. *Artif. Intell. Med.* 61, 79–88. doi: 10.1016/j.artmed.2014.03.001
- Kaur, P., Bharti, V., and Maji, S. (2020). Enhanced epileptic seizure detection using imbalanced classification. *Int. J. Recent Technol. Eng.* 9, 2412–2420. doi: 10.35940/ijrte.A2894.059120
- Khambhati, A. N., Bassett, D. S., Oommen, B. S., Chen, S. H., Lucas, T. H., Davis, K. A., et al. (2017). Recurring functional interactions predict network architecture of interictal and ictal states in neocortical epilepsy. *eNeuro* 8:ENEURO.0091-16.2017. doi: 10.1523/ENEURO.0091-16.2017
- Kiral-Kornek, I., Roy, S., Nurse, E., Mashford, B., Karoly, P., Carroll, T., et al. (2018). Epileptic seizure prediction using big data and deep learning: toward a mobile system. *EBioMedicine* 27, 103–111. doi: 10.1016/j.ebiom.2017.11.032
- Krawczyk, B., Wozniak, M., and Schaefer, G. (2014). Cost-sensitive decision tree ensembles for effective imbalanced classification. *Appl. Soft Comput.* 14, 554–562. doi: 10.1016/j.asoc.2013.08.014
- Kubat, M., and Matwin, S. (1997). “Addressing the curse of imbalanced training sets: one-sided selection,” in *Proceedings of the 14th International Conference on Machine Learning*, 179–186
- Kuncheva, L. I. (2014). *Combining Pattern Classifiers: Methods and Algorithms, 2nd Edn.* John Wiley & Sons. doi: 10.1002/9781118914564
- Lagarde, S., Roehri, N., Lambert, I., Trebuchon, A., McGonigal, A., Carron, R., et al. (2018). Interictal stereotactic-EEG functional connectivity in refractory focal epilepsies. *Brain* 141, 2966–2980. doi: 10.1093/brain/awy214
- Laurikkala, J. (2001). “Improving Identification of Difficult Small Classes by Balancing Class Distribution,” in *Artificial Intelligence in Medicine. AIME 2001. Lecture Notes in Computer Science, Vol. 2101*, eds S. Quaglini, P. Barahona, and S. Andreassen (Berlin; Heidelberg: Springer), 63–66. doi: 10.1007/3-540-48229-6_9
- Lee, P. H. (2014). Resampling methods improve the predictive power of modeling in class-imbalanced datasets. *Int. J. Environ. Res. Public Health.* 11, 9776–9789. doi: 10.3390/ijerph110909776
- Lopes da Silva, F., Pijn, J. P., and Boeijinga, P. (1989). Interdependence of EEG signals: linear vs. nonlinear associations and the significance of time delays and phase shifts. *Brain Topogr.* 2, 9–18. doi: 10.1007/BF01128839
- López, V., Fernández, A., García, S., Palade, V., and Herrera, F. (2013). An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. *Inf. Sci.* 250, 113–141. doi: 10.1016/j.ins.2013.07.007
- Loyola-González, O., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., and García-Borroto, M. (2016). Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. *Neurocomputing* 175, 935–947. doi: 10.1016/j.neucom.2015.04.120
- Lüders, H. O., Najm, I., Nair, D., Widdess-Walsh, P., and Bingman, W. (2006). The epileptogenic zone: general principles. *Epilept. Disord.* 8 (Suppl. 2):S1–9.
- Luque, A., Carrasco, A., Martín, A., and de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.* 91, 216–223. doi: 10.1016/j.patcog.2019.02.023

- Masum, M., Shahriar, H., and Haddad, H. M. (2020). "Epileptic seizure detection for imbalanced datasets using an integrated machine learning approach," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS (Montréal, QC)*, 5416–5419. doi: 10.1109/EMBC44109.2020.9175632
- Mena, L., and Gonzalez, J. A. (2006). "Machine learning for imbalanced datasets: application in medical diagnostic," in *FLAIRS 2006—Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference (Melbourne Beach, FL)*, 574–579.
- Narasimhan, S., Kundassery, K. B., Gupta, K., Johnson, G. W., Wills, K. E., Goodale, S. E., et al. (2020). Seizure-onset regions demonstrate high inward directed connectivity during resting-state: an SEEG study in focal epilepsy. *Epilepsia* 61, 2534–2544. doi: 10.1111/epi.16686
- Oldham, S., Fulcher, B., Parkes, L., Arnatkeviciute, A., Suo, C., and Fornito, A. (2019). Consistency and differences between centrality measures across distinct classes of networks. *PLoS ONE* 14:e0220061. doi: 10.1371/journal.pone.0220061
- Olejarczyk, E., Marzetti, L., Pizzella, V., and Zappasodi, F. (2017). Comparison of connectivity analyses for resting state EEG data. *J. Neural Eng.* 14, 1–13. doi: 10.1088/1741-2552/aa6401
- Peter, J., Khosravi, M., Werner, T. J., and Alavi, A. (2018). Global temporal lobe asymmetry as a semi-quantitative imaging biomarker for temporal lobe epilepsy lateralization: a machine learning classification study. *Hell. J. Nucl. Med.* 21, 95–101. doi: 10.1967/s002449910800
- Rokach, L. (2010). Ensemble-based classifiers. *Artif. Intell. Rev.* 33, 1–39. doi: 10.1007/s10462-009-9124-7
- Roland, J. L., Griffin, N., Hacker, C. D., Vellimanna, A. K., Akbari, S. H., Shimony, J. S., et al. (2017). Resting-state functional magnetic resonance imaging for surgical planning in pediatric patients: a preliminary experience. *J. Neurosurg. Pediatr.* 20, 583–590. doi: 10.3171/2017.6.PEDS1711
- Romaissa, D., Habib, M., and Chikh, M. A. (2019). "Epileptic seizure detection from imbalanced EEG signal," in *2019 International Conference on Advanced Electrical Engineering, ICAEE 2019 (Macau)*, 1–6. doi: 10.1109/ICAEE47123.2019.9015113
- Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52, 1059–1069. doi: 10.1016/j.neuroimage.2009.10.003
- Ryvlin, P., Cross, J. H., and Rheims, S. (2014). Epilepsy surgery in children and adults. *Lancet Neurol.* 13, 1114–1126. doi: 10.1016/S1474-4422(14)70156-5
- Seiffert, C., Khoshgoftar, T. M., Van Hulse, J., and Napolitano, A. (2010). RUSBoost: a hybrid approach to alleviating class imbalance. *IEEE Trans. Syst. Man, Cybern. Part A Syst. Hum.* 40, 185–197. doi: 10.1109/TSMCA.2009.2029559
- Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *J. Am. Stat. Assoc.* 81, 826–831. doi: 10.1080/01621459.1986.10478341
- Silfverhuth, M. J., Hintsala, H., Kortelainen, J., and Seppänen, T. (2012). Experimental comparison of connectivity measures with simulated EEG signals. *Med. Biol. Eng. Comput.* 50, 683–688. doi: 10.1007/s11517-012-0911-y
- Soriano, M. C., Niso, G., Clements, J., Ortín, S., Carrasco, S., Gudín, M., et al. (2017). Automated detection of epileptic biomarkers in resting-state interictal MEG data. *Front. Neuroinform.* 11:43. doi: 10.3389/fninf.2017.00043
- Spencer, S., and Huh, L. (2008). Outcomes of epilepsy surgery in adults and children. *Lancet Neurol.* 7, 525–537. doi: 10.1016/S1474-4422(08)70109-1
- Stefanowski, J., and Wilk, S. (2008). "Selective pre-processing of imbalanced data for improving classification performance," in *Data Warehousing and Knowledge Discovery Lecture Notes in Computer Science*, eds I. Y. Song, J. Eder, T. M. Nguyen (Berlin; Heidelberg: Springer), 5182. doi: 10.1007/978-3-540-85836-2_27
- Sun, Y., Kamel, M. S., Wong, A. K. C., and Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognit.* 12, 3358–3378. doi: 10.1016/j.patcog.2007.04.009
- Talairach, J., and Bancaud, J. (1966). Lesion, "irritative" zone and epileptogenic focus. *Confin. Neurol.* 27, 91–94. doi: 10.1159/000103937
- Tang, S., and Chen, S. P. (2008). "The generation mechanism of synthetic minority class examples," in *International Conference on Information Technology and Applications in Biomedicine (Shenzhen)*, 444–447. doi: 10.1109/ITAB.2008.4570642
- Varotto, G., Tassi, L., Franceschetti, S., Spreafico, R., and Panzica, F. (2012). Epileptogenic networks of type II focal cortical dysplasia: a stereo-EEG study. *Neuroimage* 61, 591–598. doi: 10.1016/j.neuroimage.2012.03.090
- Varotto, G., Tassi, L., Rotondi, F., Spreafico, R., Franceschetti, S., and Panzica, F. (2013). "Effective brain connectivity from intracranial eeg recordings: identification of epileptogenic zone in human focal epilepsies," in *Modern Electroencephalographic Assessment Techniques*, ed V. Sakkalis (New York, NY: Humana Press). doi: 10.1007/7657_2013_61
- Vlachos, I., Krishnan, B., Treiman, D. M., Tsakalis, K., Kugiumtzis, D., and Iasemidis, L. D. (2017). The concept of effective inflow: application to interictal localization of the epileptogenic focus from iEEG. *IEEE Trans. Biomed. Eng.* 64, 2241–2252. doi: 10.1109/TBME.2016.2633200
- Wang, S., and Yao, X. (2009). "Diversity analysis on imbalanced data sets by using ensemble models," in *2009 IEEE Symposium on Computational Intelligence and Data Mining*, 324–331. doi: 10.1109/CIDM.2009.4938667
- Weiss, G. M. (2004). Mining with rarity: a unifying framework. *ACM SIGKDD Explorat. Newsl.* 6, 7–19. doi: 10.1145/1007730.1007734
- Wendling, F., Ansari-Asl, K., Bartolomei, F., and Senhadji, L. (2009). From EEG signals to brain connectivity: a model-based evaluation of interdependence measures. *J. Neurosci. Methods* 183, 9–18. doi: 10.1016/j.jneumeth.2009.04.021
- Wendling, F., Chauvel, P., Biraben, A., and Bartolomei, F. (2010). From intracerebral EEG signals to brain connectivity: identification of epileptogenic networks in partial epilepsy. *Front. Syst. Neurosci.* 4:154. doi: 10.3389/fnsys.2010.00154
- Xie, C., Du, R., Ho, J. W., Pang, H. H., Chiu, K. W., Lee, E. Y., et al. (2020). Effect of machine learning re-sampling techniques for imbalanced datasets in 18F-FDG PET-based radiomics model on prognostication performance in cohorts of head and neck cancer patients. *Eur. J. Nucl. Med. Mol. Imaging* 47, 2826–2835. doi: 10.1007/s00259-020-04756-4
- Yen, S. J., and Lee, Y. S. (2006). "Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset," in *Intelligent Control and Automation. Lecture Notes in Control and Information Sciences*, eds D. S. Huang, K. Li, G. W. Irwin (Kunming: Springer). doi: 10.1007/978-3-540-37256-1_89
- Zhou, Z. H., and Liu, X. Y. (2006). "Training cost-sensitive neural networks with methods addressing the class imbalance problem," in *IEEE Transactions on Knowledge and Data Engineering*, 63–77. doi: 10.1109/TKDE.2006.17

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Varotto, Susi, Tassi, Gozzo, Franceschetti and Panzica. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.