Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 246 (2024) 4600–4609

Procedia Computer Science

www.elsevier.com/locate/procedia

28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

Optimizing machine learning models for classification of stroke patients with epileptiform EEG pattern: the impact of dataset balancing techniques

Katerina Iscra[a]*, Alessandro Biscontin[a,b], Aleksandar Miladinovic[a,b], Andrea Bonini[a], Giovanni Furlanis[c], Gabriele Prandin[c], Michele Malesani[c], Marcello Naccarato[c], Paolo Manganotti[c], Agostino Accardo[a], Miloš Ajčević[a]

[a]*Departement Engineering and Architecture, University of Trieste, Trieste, Italy*
[b]*Istitute for Maternal and Child Health – IRCCS Burlo Garofolo, Trieste, Italy*
[c]*Clinical Unit of Neurology, Department of Medicine, Surgery and Health Sciences, Trieste University Hospital ASUGI, Trieste, Italy*

Abstract

Epileptiform electroencephalogram (EEG) patterns are commonly observed in stroke patients and can significantly impact clinical management and patient outcomes. Therefore, the classification of the stroke patients in order to identify the subjects with high probability of epileptiform EEG patterns may improve the stroke management. In recent years, there has been a notable increase in interest and utilization of machine learning, especially in the domain of classification tasks. Nevertheless, the presence of imbalanced datasets presents hurdles for machine learning algorithms, resulting in skewed predictions toward dominant classes and diminished accuracy, especially for underrepresented ones. Hence, the study aims to evaluate the effects of dataset balancing methods on the classification efficacy of machine learning models for classification of stroke patients with epileptiform EEG patterns by conducting a comparative analysis between models trained on imbalanced and balanced datasets. Four different sampling techniques were employed: an oversampling technique, SMOTENC; an undersampling technique, NearMiss; and two techniques that combine over- and undersampling methods, SMOTEToken and SMOTEENN. The features selection was made using the ReliefF scoring method and for model construction, only features that presented a contribution value greater than 0.01 were utilized. Five different machine learning models were considered in the study: classification tree, logistic regression, naïve Bayes, artificial neural network and support vector machine. The produced models were trained on the original and resampled training set and subsequently the models' performances were evaluated on the test set. The results showed that SMOTENC was the most effective among the considered dataset balancing techniques, showing superior classification

* Katerina Iscra. Tel.: +39 040 558 7130.
E-mail address: katerina.iscra@phd.units.it

Peer-review under responsibility of the scientific committee of the 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems
10.1016/j.procs.2024.09.324

performance compared to other methods and the original dataset. Models utilizing SMOTENC exhibited significant improvements in AUC (0.76 vs 0.67) and specificity values (0.73 vs 0.43) while maintaining comparable accuracy (0.72 vs 0.74) to those trained on the original dataset, respectively. Furthermore, it has been noted that different sampling techniques result in different selection of the most predictive features. In conclusion, our study highlights the crucial role of utilizing dataset balancing methods to improve the classification performances of predictive models in case of highly unbalanced datasets such as case of stratification of stroke patients with epileptiform EEG patterns.

Peer-review under responsibility of the scientific committee of the 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems

Keywords: Dataset balancing techniques, Classification Models, Stroke, Epileptiform EEG patterns

1. Introduction

Epileptiform electroencephalogram (EEG) patterns are commonly observed in stroke patients and can significantly impact clinical management and patient outcomes [1]. Stroke is the second most common cause of death and the third most common cause of disability worldwide [2]. According to the World Health Organization (WHO), around 17 million people worldwide suffer from strokes each year. There are different types of strokes, with ischemic and hemorrhagic stroke being the two main types.

Accurate detection and classification of EEG patterns are crucial for timely intervention and treatment planning[2, 3]. Therefore, the classification of the stroke patients in order to identify the subjects with high probability of epileptiform EEG patterns may improve the stroke management. EEG offers invaluable real-time and dynamic insights that can significantly enhance prognostic accuracy [4]. Moreover, it stands out as the foremost neurodiagnostic tool for identifying epileptic activity, especially in cases involving non-convulsive post-stroke epileptic discharges [4, 5].

In recent years, the field of machine learning, particularly in the realm of classification tasks, has witnessed a surge in interest and application. Some studies have developed prediction models for post-stroke outcomes [6–8], which could aid in clinical decision-making and the management of post-stroke patients. In particular, the SeLECT score, based on five clinical predictors, showed good predictive ability in three external validation cohorts, offering potential for identifying individuals at high risk of seizures and guiding personalized medicine approaches, such as selecting candidates for antiepileptogenic treatment trials and biomarker studies of epileptogenesis [8].

One critical aspect that often determines the efficacy of these classifiers is the quality and balance of the dataset used for training. The balance of a dataset refers to the distribution of samples across different classes or categories. Real-world data tends to be unbalanced, particularly in medical datasets where high-risk patients are often outnumbered by the low-risk group [9]. Imbalance of the data-set is a primary factor contributing to decreased generalization in machine learning algorithms [10] and presents challenges in constructing effective classifiers [11]. Indeed, the classifiers built on the basis of an unbalanced dataset tend to bias their predictions towards the majority classes, often ignoring the minority ones. As a result, the performance of classifiers can be severely impacted, leading to suboptimal results, particularly for the minority classes.

Therefore, addressing this issue often necessitates the utilization of effective sampling techniques in medical datasets, under-sampling the majority of samples and/or oversampling minority classes to achieve a balanced distribution of sample classes [12]. Under-sampling involves removing some data from the majority class, while oversampling entails increasing the minority class by adding artificially generated or duplicated data [13]. Understanding the impact of dataset balancing techniques on classification performance is crucial for developing robust and reliable machine learning models.

Therefore, the aim of the study is to investigate the impact of dataset balancing strategies on the classification performance of machine learning models for classification of stroke patients with epileptiform EEG patterns. We compare the performance of models trained on imbalanced datasets against those trained on balanced datasets using various balancing techniques.

2. Methods

2.1. Study population and protocol

The study encompassed 455 consecutive stroke patients including both ischemic and hemorrhagic stroke admitted at stroke unit of neurological clinic of Trieste University Hospital, 349 (149F/200M, aged: 75 ± 14) without epileptiform EEG patterns observed and 106 (66F/40M, aged: 76 ± 12) with epileptiform EEG patterns detected. All included patients underwent common neurologic stroke workup, stroke risk factors assessment, neurological examination at admission including National Institutes of Health Stroke Scale (NIHSS), as well as neuroimaging assessment at admission and 24h follow-up. All patients underwent EEG at least two recordings during the first week from admission and the epileptiform EEG patterns were assessed by three experienced neurologists, in order to define the two classification groups. The following set of features to be considered for the development of classification models were extracted: Sex, Age, NIHSS at admission, Hypercholesterolemia, Atrial fibrillation, Diabetes, Heart failure, Smoke, Renal failure, Hypertension, Coronary disease, Previous cognitive impairment, Acute infectious state, Previous ischemic or hemorrhagic stroke, Type of stroke (ischemic or hemorrhagic), Cortex lesion, Supratentorial stroke, Number of interested lobes, Rreperfusion treatment

2.2. Sampling Techniques

Before applying the sampling techniques, a test set corresponding to 10% of the original dataset was randomly extracted. Four different sampling techniques were then employed on the training set, composed by the remaining 90% of the original dataset: an oversampling technique, SMOTENC; an undersampling technique, NearMiss; and two techniques that combine over- and undersampling methods, SMOTEToken and SMOTEENN.

SMOTENC - Synthetic Minority Over-sampling Technique for Nominal and Continuous [14]. SMOTENC is a version of the SMOTE algorithm that considers datasets containing both numeric and categorical characteristics. This oversampling approach, just like SMOTE, increases the representation of the minority class by generating "synthetic" instances in a way that is less tailored to specific applications, working within "feature space" rather than "data space". The minority class is upsampled by incorporating synthetic instances along the line segments connecting the nearest neighbors of the minority class. Synthetic samples are produced by calculating the difference between the feature vector (sample) and its nearest neighbor, then multiplying this difference by a random number ranging from 0 to 1, and finally adding the resulting value to the feature vector. This process selects a random point along the line segment between two specific elements. Consequently, this method broadens the decision region of the minority class, , enhancing its overall generality. Thus, synthetic examples encourage the classifier to create larger and less specific decision regions.

NearMiss [15]. NearMiss is a subsampling technique. Specifically, it encompasses three distinct NearMiss algorithms, each relying on some heuristic rules for subsampling. NearMiss-1 identifies samples from the majority class where the average distance to the k' nearest samples of the minority class is minimized. In contrast, NearMiss-2 identifies samples of the majority class where the average distance to the farthest samples of the minority class is minimized. The third method, NearMiss-3, operates in two stages: initially, it retains the m nearest neighbors for each minority sample; subsequently, it selects majority samples based on their maximum average distance to the k nearest neighbors. When conducting under-sampling on a particular class, the NearMiss-1 method may be influenced by noise. NearMiss-2, on the other hand, remains impervious to noise as it prioritizes the farthest samples over the nearest ones, although the presence of noise could potentially influence the sampling process, particularly in the context of marginal outliers. NearMiss-3 is likely to be the least impacted by noise, primarily because of its initial step in sample selection. Also for this reason, the third algorithm, NearMiss-3, was used in the study.

SMOTETomek [16]. Applying only over- or undersampling sometimes does not lead to an optimal result. Frequently, class boundaries are ambiguous, with instances of predominant classes encroaching upon minority class territories, while minority class interpolations might extend the boundaries of the minority group, inadvertently inserting artificial instances into the majority class domain. Training a classifier under such circumstances can result in overfitting. Improving the definition of class clusters can be achieved by employing Tomek links as a data cleaning technique on the oversampled training set using SMOTE. Tomek links [17] can be defined as follows:

considering Ei and Ej two examples from distinct classes, and d(Ei,Ej) like the distance between them, a pair (Ei,Ej) constitutes a Tomek link if there no exists another example, El, such that the distance between Ei and El is less than d(Ei,Ej), or the distance between Ej and El is less than d(Ei,Ej). In the case of SMOTEToken, rather of only removing instances from the majority classes forming Tomek links, istances from both classes are removed. In particular, the initial dataset is upsampled through SMOTE, followed by the identification and removal of Tomek links, resulting in a balanced dataset characterized by distinct class clusters.

SMOTEENN [16]. This method, similar to SMOTEToken, combines over- and under-sampling using SMOTE and another undersampling technique, Edited Nearest Neighbours (ENN) [18]. The ENN algorithm cleans the dataset by removing examples situated near the decision boundary, discarding any example whose class label differs from the class of at least two of its three nearest neighbors. This method involves training a K-Nearest Neighbors model on the entire dataset to identify each observation's K closest neighbors, particularly focusing on the targeted classes. It then proceeds to remove observations if the majority of their neighbors belong to a different class, thus purifying the dataset by eliminating instances near the decision boundary. Therefore, the ENN method retains instances from the majority class when the majority, or all, of their neighbors share the same class. ENN typically removes a greater number of instances compared to Tomek links, offering a more comprehensive data cleaning approach. Similar to what was done with SMOTEToken, the initial dataset is upsampled with SMOTE and then cleaned with the ENN algorithm, ultimately providing a dataset with two well-balanced classes.

By applying abovementioned techniques, four new datasets were created. In particular, with the SMOTENC oversampling technique, 219 subjects with epileptic patterns in the EEG were added. Through the NearMiss subsampling technique, 219 patients without epileptic patterns in the EEG were excluded from the analysis. The SMOTETomek technique added 167 subjects with epileptic patterns and eliminated 52 subjects without epileptic patterns. Finally, with SMOTEEEN, 121 patients with epileptic patterns in the EEG were created and 98 patients without these patterns were excluded. These datasets were subsequently used to produce the models based on the machine learning techniques described in the following section.

2.3. Feature selection and classification

The features selection was made using the ReliefF scoring method [19], a proficient heuristic estimator. ReliefF is adept at handling datasets with both conditionally dependent and independent attributes, as well as noisy, incomplete, and multi-class datasets, while assessing an attribute's ability to distinguish between two classes on similar data instances. In particular, it searches out k nearest hits or misses rather than just one, and then calculates the average contribution of all these k nearest hits or misses. For model construction, only features that presented a contribution value greater than 0.01 were utilized.

Five different machine learning models were considered in the study: classification tree, logistic regression, naïve Bayes, artificial neural network and support vector machine.

The classification tree [20], represents one of the simplest methods and consists of two main phases: tree growth and pruning. Initially, the tree expands by selecting splits that result in nodes containing only one component class. Through the mean square error (MSE) it is possible to evaluate the impurity of each node during the growth of the tree. Subsequently, the tree is pruned using a minimum cost complexity function, considering both the number of nodes and the potential for misclassification. The result of this phase is the identification of the best subtree that minimizes the cost-complexity function. The constructed decision tree is a binary tree with the minimum number of instances in the leaves being 2. It does not further subdivide subsets that contain less than 5 instances and has a maximum depth of 100 levels. Additionally, the tree stops splitting nodes when it reaches a 95% majority threshold.

Another machine learning algorithm is based on logistic regression. Typically, the sigmoid function within a logistic regression classifier linearly combines multiple feature values or explanatory variables. The sigmoid function produces outputs ranging from 0 to 1, with the midpoint serving as the threshold for class assignment. Inputs that produce results above 0.5 are classified into class 1, while those below 0.5 are assigned to class 0 [21]. The logistic regression classification algorithm used in the study employs ridge regularization with a cost strength set to 1.

A naive Bayesian classifier, based on Bayesian statistics, employs strong independence assumptions to classify data. Despite its simplicity, the naive Bayes classifier performs exceptionally well and often outperforms more complex algorithms [22]. One of its strengths is that it requires only a modest amount of training data to estimate classification parameters, which can prove highly beneficial. In the study, the model based on the Naive Bayes

algorithm applies data preprocessing. This involves the elimination of any empty columns, if there are any, and the division of numerical values into 4 bins of uniform frequency.

Artificial neural networks are computational models that are based on and imitate the function of the human brain [23]. They are composed of a large number of 'neurons', which are processing nodes connected to each other by weighted connections. Feed-forward networks, a common type of neural network [23], are composed of three layers: the input layer, where the problem inputs are received; the hidden layers, where each neuron applies an activation function to the weighted sum of the input values and transmits its output to subsequent nodes; and an output layer, which outputs the final results of the problem. A neural network is constructed through a learning process, which can be supervised, unsupervised, or reinforced. Throughout this process, network parameters, such as weights and biases, are iteratively adjusted through interaction with the training data or the environment in which the network operates [23]. Thus, the process of adapting the parameters of the neural network is what enables the network to enhance its performance in the specific task for which it was designed. In the study, a neural network with a hidden layer consisting of 100 neurons was constructed, using the ReLU (Rectified Linear Unit) activation function and a gradient-based stochastic optimizer for weight optimization. The maximum number of iterations used was 200.

The Support Vector Machine (SVM) is a tool capable of both classification and prediction via regression [24]. Using machine learning theory, SVM seeks to maximize predictive accuracy by automatically avoiding overfitting to input data. Its formulation is based on the principle of structural risk minimization (SRM), giving it greater generalization capacity. Classification with SVM is an example of supervised learning, where known labels help the system learn to perform correctly and validate the accuracy of predictions. In the study, the SVM was configured to minimize the error function with a cost equal to 1. As regards the kernel, i.e. the mathematical function used to transform the data into a higher dimensional space where a separating hyperplane between the classes, the Radial Basis Function (RBF) was selected. The number of iterations was limited to 100.

2.4. Evaluation of sampling techniques and models

The produced models were trained on the resampled training sets as well as on the original training set. Subsequently, the classification performance of the produced models was evaluated by the following metrics on priviously unseen data (test set): the classification accuracy, the area under the curve (AUC), F1 measure, precision, recall, and specificity. Moreover, the confusion matrix and ROC analysis were executed for each model. All analysis was conducted in Python Orange3 Data Mining library and toolbox [25].

3. Results

In Table 1 the performances of the produced models considering the original training set, without application of any of balancing methods are reported.

Table 1. Performance measures of the models produced considering original training set and evaluated by the test set.

Models	CA	AUC	F1	Precision	Recall	Specificity
Classification tree	0.69	0.54	0.69	0.68	0.69	0.40
Logistic regression	0.77	0.67	0.71	0.72	0.77	0.30
Naïve Bayes	0.74	0.67	0.72	0.72	0.74	0.43
Neural Network	0.75	0.62	0.72	0.71	0.75	0.39
SVM	0.76	0.60	0.71	0.70	0.76	0.31

Although the overall accuracy values were moderately high for the specific clinical research question, the AUC and the specificity were very low. The underperformance observed for the AUC and specificity indicates that the models were affected by the fact that the two classes are not balanced; in fact, subjects with epileptiform graphs are much fewer than those without.

Table 2 presents the performance of the produced models considering the different under- and oversampling balancing techniques (SMOTENC, NearMiss, SMOTETomek, SMOTEENN).

Table 2. Performance measures of the models produced considering under- and oversampling balancing techniques SMOTENC, NearMiss, SMOTETomek, SMOTEENN and evaluated by the test set.

	Models	CA	AUC	F1	Precision	Recall	Specificity
SMOTENC							
	Classification tree	0.60	0.46	0.62	0.63	0.60	0.37
	Logistic regression	0.72	0.75	0.73	0.74	0.72	0.54
	Naïve Bayes	0.72	0.76	0.74	0.79	0.72	0.73
	Neural Network	0.74	0.77	0.75	0.79	0.74	0.69
	SVM	0.65	0.53	0.67	0.70	0.65	0.52
NearMiss							
	Classification tree	0.49	0.53	0.53	0.65	0.49	0.52
	Logistic regression	0.60	0.64	0.63	0.71	0.60	0.60
	Naïve Bayes	0.49	0.48	0.53	0.63	0.49	0.47
	Neural Network	0.57	0.63	0.60	0.72	0.57	0.64
	SVM	0.62	0.68	0.65	0.76	0.62	0.70
SMOTETomek							
	Classification tree	0.62	0.44	0.61	0.61	0.62	0.23
	Logistic regression	0.65	0.56	0.67	0.70	0.65	0.52
	Naïve Bayes	0.63	0.60	0.65	0.68	0.63	0.47
	Neural Network	0.63	0.63	0.65	0.69	0.63	0.52
	SVM	0.66	0.52	0.65	0.65	0.66	0.34
SMOTEENN							
	Classification tree	0.43	0.46	0.47	0.64	0.43	0.55
	Logistic regression	0.48	0.59	0.51	0.66	0.48	0.56
	Naïve Bayes	0.55	0.66	0.58	0.76	0.55	0.73
	Neural Network	0.57	0.66	0.60	0.74	0.57	0.68
	SVM	0.63	0.69	0.66	0.76	0.63	0.70

Among the considered dataset balanced techniques, SMOTENC showed better results compared to other techniques and original dataset. In particular, the models utilizing the SMOTEC balancing technique demonstrated superior performance compared to those employing other balancing methods. While the accuracy values remained comparable to models trained on the original dataset, notable improvements were observed in terms of AUC and specificity values. Figure 1 represents the ROC curves of all the models built using the training set resampled with SMOTENC technique.

By observing the ROC curves and classification performance reported in Table 2, the naive Bayes presented the best performance when considering both overall accuracy which was comparable to neural network and logistic regression, while showing the notably better performance in terms of specificity compared to logistic regression and slightly better specificity then neural network. In Figure 2 the comparison between ROC curves for the two best models (naïve Bayes and neural network) against the results obtained from models produced the same classification techniques on the original training set was reported.

From the Fig. 2 it can be observed that the models constructed from the resampled training set using the oversampling technique outperform those obtained from the unbalanced training set. Furthermore, when exclusively considering the models trained on the original dataset, we observe that, once again, the Naive Bayes model demonstrates slight classification superiority. To support the previous findings, in Table 3 we report also a comparative analysis of the confusion matrices for Naive Bayes models constructed using the original dataset and another employing SMOTENC resampled data.

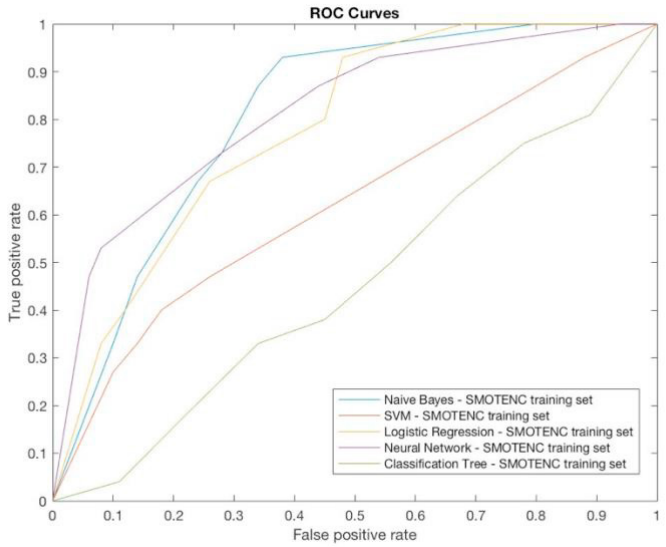

Fig. 1. ROC curves of all models built using the training set resampled with SMOTENC technique

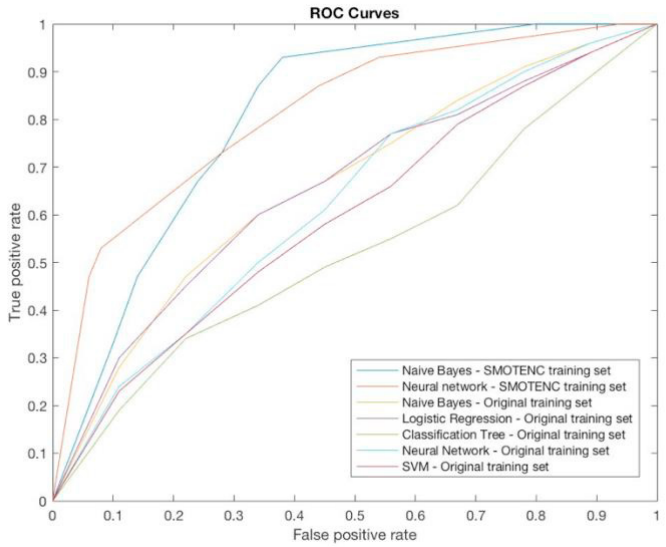

Fig. 2. ROC curves for the two best models (naïve Bayes and neural network) against the results obtained from models produced the same classification techniques on the original training set

Table 3. Confusion matrices obtained by (a) Naïve Bayes model constructed with training set resampled with SMOTENC technique; (b) Naïve Bayes model constructed with original training set.

a)	noEPI	EPI	Σ	b)	noEPI	EPI	Σ
noEPI	72%	28%	35	noEPI	87%	13%	35
EPI	27%	73%	11	EPI	70%	30%	11
Σ	28	18	46	Σ	38	8	46

Table 4 shows the selected features utilizing the ReliefF method as the metric for each training sets derived from different sampling techniques. The data reported in the table shows how various sampling techniques implicate different set of most significant features.

Table 4. The set of selected features for each sampling technique.

Training dataset	Features
Original	Reperfusion treatment, Hypercholesterolemia, Cortex lesion, Sex, Supratentorial stroke, NIHSS at admission, Diabetes, Smoke, Acute infectious state, Number of interested lobes, Type of stroke (ischemic or hemorrhagic), Renal failure, Age, Previous ischemic or hemorrhagic stroke, Coronary disease
SMOTENC	Sex, Type of stroke (ischemic or hemorrhagic), Age, NIHSS at admission, Atrial fibrillation, Diabetes, Supratentorial stroke, Hypercholesterolemia, Reperfusion treatment, Cortex lesion, Hypertension
NearMiss	Sex, Hypertension, NIHSS at admission, Acute infectious state, Supratentorial stroke, Age, Previous cognitive impairment, Reperfusion treatment
SMOTETomek	NIHSS at admission, Age, Hypercholesterolemia, Reperfusion treatment, Type of stroke (ischemic or hemorrhagic), Number of interested lobes, Supratentorial stroke, Diabetes, Hypertension, Cortex lesion, Heart failure, Coronary disease
SMOTEENN	NIHSS at admission, Hypercholesterolemia, Reperfusion treatment, Cortex lesion, Sex, Age, Number of interested lobes, Diabetes, Type of stroke (ischemic or hemorrhagic), Acute infectious state, Smoke, Previous ischemic or hemorrhagic stroke, Renal failure, Coronary disease

4. Discussion

The classification of the stroke patients in order to identify the subjects with high probability of epileptiform EEG patterns may improve the stroke management. In this study we aimed to assess the impact of dataset balancing techniques on the classification performance of machine learning models for identification of stroke patients with epileptiform EEG patterns by conducting a comparative analysis between models trained on imbalanced and balanced datasets. The study results showed that the SMOTENC oversampling technique may improve the classification performance also for the minority classes such as stroke patients with epileptiform EEG patterns. In particular, we observed the improvement of the AUC and specificity while manteining the overall classification accuracy compared to the original dataset. In addition, in this preliminary study on the limited dataset the model produced with naive Bayes approach showed the best classification performance amoung others on SMOTENC oversampling dataset.

Several studies have demonstrated that oversampling techniques yield superior results [26–28]. In particular, Alahmari investigated the impact of class imbalance on classification models using a real dataset concerning the screening of autism spectrum disorders (ASD) [26]. His study revealed the advantages of oversampling techniques, showcasing improvements in specificity, sensitivity, and precision during the pre-processing phase. Their empirical findings, derived from experiments employing Random Forest and Naïve Bayes classifiers, underscored the severity of class imbalance as a potential source of bias. Notably, models generated by Random Forest and Naïve Bayes exhibited enhanced performance metrics—including ROC, sensitivity, specificity, and accuracy rates—when the dataset underwent resampling via random oversampling and SMOTE techniques. Mohamed et al. also demonstrated that oversampling works better than undersampling for different classifiers and achieves higher scores in various evaluation metrics [27]. Khushi et al. conducted a comprehensive study on class imbalance techniques for predicting lung cancer presence [28]. Their investigation encompassed various data-level and hybrid systems. They compared 23 unbalanced learning methods, comprising ten undersampling techniques, seven oversampling techniques, two

hybrid resampling methods, and four hybrid systems. Among these methods, they found that oversampling yielded the most promising results. Indeed, most models employing oversampling demonstrated higher AUC values compared to others. Notably, random forest with random oversampling emerged as the top predictive model. Conversely, akin to our findings, models utilizing the Near Miss subsampling technique exhibited inferior performance.

The results derived from models trained with the training set balanced using the SMOTENC oversampling technique outperform those obtained from models produced on the original training set. While accuracy values were higher then 70%, specificity and AUC values did not exhibit the same level of performance. This discrepancy is further highlighted by the ROC curves, which notably lag behind those generated by the Naive Bayes and Neural Network models trained with the SMOTENC-enhanced dataset. The obtained confusion matrices also indicate that the Naive Bayes model trained with the SMOTENC training set outperforms those trained with the original training set, particularly in classifying the minority group, namely stroke subjects with epileptiform EEG patterns. The underlying cause is undoubtedly the class imbalance inherent in the original dataset. Studies have consistently shown that machine learning classifiers trained on imbalanced datasets tend to exhibit bias towards the majority classes while neglecting the minority ones [29, 30]. This imbalance can significantly impair classifier performance, leading to suboptimal results, particularly for the underrepresented classes. Consequently, it is imperative to address this issue by adopting techniques such as under- or oversampling. These methods aim to balance the class distribution within the dataset, thus mitigating the bias towards the majority classes and improving the classifier's ability to accurately predict outcomes across all classes. Therefore, it is crucial for researches to not only implement these sampling techniques but also to comprehend their impact on the classification performance of machine learning models, as we have thoroughly investigated in our study. Moreover, in future work, deep learning techniques for data augmentation [31, 32] can also be considered if a substantial number of original samples are available for unbalanced datasets.

Finally, the study results reveal that different balancing techniques yield distinct selections of features for model construction. Specifically, employing the SMOTENC oversampling technique the following features were selected: Sex, Type of stroke (ischemic or hemorrhagic), Age, NIHSS at admission, Atrial fibrillation, Diabetes, Supratentorial stroke, Hypercholesterolemia, Reperfusion treatment, Cortex lesion, Hypertension. Notably, these selected features hold clinical significance, as they can effectively gauge the severity of stroke, often associated with complications such as epileptic EEG patterns. This precise feature selection further underscores the efficacy of the oversampling technique.

In conclusion, our study underscores the critical importance of employing dataset balancing techniques, particularly SMOTENC oversampling, to enhance the classification performance of machine learning models in classification of minority classes such as the stroke patients with epileptiform EEG patterns. The superior performance of models trained on balanced datasets, highlighted by the Naive Bayes model constructed using SMOTENC oversampling, not only emphasizes the efficacy of these techniques but also underscores their potential to improve clinical decision-making by accurately identifying crucial features associated with stroke severity and complications, such as epileptic EEG patterns.

References

[1] Bentes C, Martins H, Peralta AR, Morgado C, Casimiro C, Franco AC, Fonseca AC, Geraldes R, Canhão P, Pinho E Melo T, Paiva T, Ferro JM (2018) Early EEG predicts poststroke epilepsy. Epilepsia Open 3:203–212. https://doi.org/10.1002/epi4.12103

[2] Doerrfuss JI, Kilic T, Ahmadi M, Holtkamp M, Weber JE (2020) Quantitative and Qualitative EEG as a Prediction Tool for Outcome and Complications in Acute Stroke Patients. Clin EEG Neurosci 51:121–129. https://doi.org/10.1177/1550059419875916

[3] Lima FO, Ricardo JAG, Coan AC, Soriano DC, Avelar WM, Min LL (2017) Electroencephalography Patterns and Prognosis in Acute Ischemic Stroke. Cerebrovasc Dis 44:128–134. https://doi.org/10.1159/000477674

[4] Jordan KG (2004) Emergency EEG and continuous EEG monitoring in acute ischemic stroke. J Clin Neurophysiol 21:341–352

[5] Mecarelli O, Pro S, Randi F, Dispenza S, Correnti A, Pulitano P, Vanacore N, Vicenzini E, Toni D (2011) EEG patterns and epileptic seizures in acute phase stroke. Cerebrovasc Dis 31:191–198. https://doi.org/10.1159/000321872

[6] Galovic M, Stauber AJ, Leisi N, Krammer W, Brugger F, Vehoff J, Balcerak P, Müller A, Müller M, Rosenfeld J, Polymeris A, Thilemann S, De Marchis GM, Niemann T, Leifke M, Lyrer P, Saladin P, Kahles T, Nedeltchev K, Sarikaya H, Jung S, Fischer U, Manno C, Cereda CW, Sander JW, Tettenborn B, Weder BJ, Stoeckli SJ, Arnold M, Kägi G (2019) Development and Validation of a Prognostic Model of

Swallowing Recovery and Enteral Tube Feeding After Ischemic Stroke. JAMA Neurol 76:561–570. https://doi.org/10.1001/jamaneurol.2018.4858

[7] Chi N-F, Kuan Y-C, Huang Y-H, Chan L, Hu C-J, Liu H-Y, Chiou H-Y, Chien L-N (2018) Development and validation of risk score to estimate 1-year late poststroke epilepsy risk in ischemic stroke patients. Clinical Epidemiology 10:1001–1011. https://doi.org/10.2147/CLEP.S168169

[8] Galovic M, Döhler N, Erdélyi-Canavese B, Felbecker A, Siebel P, Conrad J, Evers S, Winklehner M, von Oertzen TJ, Haring H-P, Serafini A, Gregoraci G, Valente M, Janes F, Gigli GL, Keezer MR, Duncan JS, Sander JW, Koepp MJ, Tettenborn B (2018) Prediction of late seizures after ischaemic stroke with a novel prognostic model (the SeLECT score): a multivariable prediction model development and validation study. Lancet Neurol 17:143–152. https://doi.org/10.1016/S1474-4422(17)30404-0

[9] M. Mostafizur Rahman and D. N. Davis, Addressing the Class Imbalance Problem in Medical Datasets, International Journal of Machine Learning and Computing, Vol. 3, No. 2, April 2013

[10] M. S. Kim, "An Effective Under-Sampling Method for Class. Imbalance Data Problem," in Proc. 8th International Symposium on Advance intelligent System (ISIS 2007), 2007.

[11] Cluster-based majority under-sampling approaches for class imbalance learning | IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/5609385. Accessed 13 May 2024

[12] Yang H, Li X, Cao H, Cui Y, Luo Y, Liu J, Zhang Y (2021) Using machine learning methods to predict hepatic encephalopathy in cirrhotic patients with unbalanced data. Comput Methods Programs Biomed 211:106420. https://doi.org/10.1016/j.cmpb.2021.106420

[13] Laza R, Pavón R, Reboiro-Jato M, Fdez-Riverola F (2011) Evaluating the effect of unbalanced data in biomedical document classification. J Integr Bioinform 8:177. https://doi.org/10.2390/biecoll-jib-2011-177

[14] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic Minority Over-sampling Technique. jair 16:321–357. https://doi.org/10.1613/jair.953

[15] Kasem A, Ammar Ghaibeh A, Moriguchi H (2017) Empirical Study of Sampling Methods for Classification in Imbalanced Clinical Datasets. In: Phon-Amnuaisuk S, Au T-W, Omar S (eds) Computational Intelligence in Information Systems. Springer International Publishing, Cham, pp 152–162

[16] Batista GEAPA, Prati RC, Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor Newsl 6:20–29. https://doi.org/10.1145/1007730.1007735

[17] Provost F, Weiss GM (2003) Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. jair 19:315–354. https://doi.org/10.1613/jair.1199

[18] Wilson DL (1972) Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE Transactions on Systems, Man, and Cybernetics SMC-2:408–421. https://doi.org/10.1109/TSMC.1972.4309137

[19] Kononenko I, Šimec E, Robnik-Sikonja M (1997) Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. Applied Intelligence 7:39–55. https://doi.org/10.1023/A:1008280620621

[20] Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification And Regression Trees. Routledge, Boca Raton

[21] Urso A, Fiannaca A, La Rosa M, Ravì V, Rizzo R (2019) Data Mining: Prediction Methods. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C (eds) Encyclopedia of Bioinformatics and Computational Biology. Academic Press, Oxford, pp 413–430

[22] Rish I (2001) An empirical study of the naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. pp 41–46

[23] A.D.Dongare, R.R.Kharde, D.Kachare A (2012) Introduction to Artificial Neural Network

[24] Jakkula, Vikramaditya. "Tutorial on support vector machine (svm)." School of EECS, Washington State University 37.2.5 (2006): 3.

[25] Demšar J, Curk T, Erjavec A, Demsar J, Curk T, Erjave A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B Orange: Data Mining Toolbox in Python. 5

[26] Alahmari F (2020) A Comparison of Resampling Techniques for Medical Data Using Machine Learning. J Info Know Mgmt 19:2040016. https://doi.org/10.1142/S021964922040016X

[27] Mohammed R, Rawashdeh J, Abdullah M (2020) Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. In: 2020 11th International Conference on Information and Communication Systems (ICICS). pp 243–248

[28] Khushi M, Shaukat K, Alam TM, Hameed IA, Uddin S, Luo S, Yang X, Reyes MC (2021) A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. IEEE Access 9:109960–109975. https://doi.org/10.1109/ACCESS.2021.3102399

[29] Ganganwar V (2012) An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering 2:42–47

[30] Ali H, Salleh MNM, Saedudin R, Hussain K, Mushtaq MF (2019) Imbalance class problems in data mining: a review. Indonesian Journal of Electrical Engineering and Computer Science 14:1552–1563. https://doi.org/10.11591/ijeecs.v14.i3.pp1552-1563

[31] Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y (2018) Data Synthesis based on Generative Adversarial Networks. Proc VLDB Endow 11:1071–1083. https://doi.org/10.14778/3231751.3231757

[32] Apellániz PA, Parras J, Zazo S (2024) An improved tabular data generator with VAE-GMM integration