Instance retrieval from non-labeled data as a strategy for automatic classification of imbalanced e-mail datasets

Juan Manuel Fernández[1], Marcelo Errecalde[2]

[1] Department of Basic Sciences, National University of Lujan, Argentina
[2] LIDIC, National University of San Luis, Argentina
jmfernandez@unlu.edu.ar, merreca@unsl.edu.ar

Abstract. One of the main challenges in automatic email classification problems occurs when it is necessary to work with a relatively large number of classes and the classes are highly imbalanced. That happens even when non-labeled textual bases are available because manual labeling is costly. In this respect, all automatic text classification strategies –to a greater or lesser extent– are sensitive to the problems of imbalance between classes.
The most widely used approaches for learning from unbalanced databases consists of resampling techniques, either by undersampling or oversampling the datasets. However, existing techniques have some problems to be solved.
In this work we present a new proposal that consists of balancing the classes of the data set by retrieving unlabeled instances (e-mails) that are similar to those of the minority classes. It is shown that, for the data set used, it is a valid, viable and competitive strategy with respect to the resampling strategies currently used to learn from imbalanced email databases.

Keywords: imbalanced data, automatic classification, information retrieval

1 Introduction

Text analysis and processing techniques face very complex problems within the area of computer science, mainly due to the difficulty of language analysis. That is caused by its ambiguity, mainly in the semantic analysis stage, and the relatively scarce training materials and the computational capacity required for the analysis to run certain algorithms very demanding in hardware resources. [29][5]. In particular, emails have specific characteristics concerning other textual elements that present some differences and problems between traditional text mining and *email mining.*

Regarding the problem of automatic classification of emails, it consists of assigning an email to a set of automatically predefined classes using, in general, a machine learning technique. The classification is generally performed on the basis of relevant words or features extracted from the e-mail text and, since the classes are predefined and training instances are class-labeled, it is usually addressed as a supervised machine learning task [11].

Approaches to email classification include neural networks [1], techniques based on support vector machines, Naive Bayes and TF-IDF classifiers [28], among others. More recently, Deep Learning-based approaches like *Long-Short-Term-Memory* are gaining attention to classify spam emails [10].

Finally, as an evolution in the previous strategy, in 2017, a new neural network architecture, simple and parallelizable, called *Transformer* [30] was proposed. It is exclusively based on attention mechanisms and completely dispenses with recurrence and convolutions. From these ideas arise what is known in the literature like the current state of the art of language representation models, called BERT (Bidirectional Encoder Representations from Transformers) [12]. There are uncountable studies on text classification with this representation model and, in email classifications, it has shown improvements in performance compared to previous strategies [15].

All of those automatic classification strategies - to a greater or lesser extent - are affected by class imbalance problems. Class imbalance is present in many real-world classification datasets and consists of a disproportion in the number of examples of different classes in the problem. This situation hampers the performance of classifiers due to their accuracy-oriented design, which generally results in the minority class being overlooked [14].

In this work, different well-known strategies for learning from imbalanced data are evaluated and compared against a new one, in the specific domain of e-mail classification. This proposal consists in using the set of manually labeled data that constitute the (imbalanced) training set, to select the most representative words of each minority class and then using them to retrieve new instances from a repository of unlabeled data to balance the training dataset.

The rest of the article is organized as follows. Section II presents some related works, the addressed research gap, and our working hypothesis. Section III presents the research methodology with its involved tasks and Section IV describes the experimental study and the analysis of the results. Finally, Section V gives some conclusions, contributions of our work, and possible future work.

2 Background

Most machine learning algorithms work best with balanced datasets but the problem arises when the given datasets are highly imbalanced in nature [26]. Classification of these imbalanced datasets is a complex task for traditional classifiers, as they generally tend to favor the samples of the majority classes over the minority ones. A large number of techniques have been developed [21] [25] to correctly distinguish the minority classes. These techniques can be categorized into four main groups, depending on how they deal with the problem [14]:

- Algorithm level approaches (also called internal): try to adapt existing classifier learning algorithms to bias the learning toward the minority class.

- Data level (or external) approaches: aim at rebalancing the class distribution by resampling the data space.
- Cost-sensitive approaches: allow the definition of costs associated with each of the classes in order to generate a weighting in the classification.
- Ensemble-based methods: usually consist of a combination of an ensemble learning algorithm and one of the above techniques.

One of the most widely used is the data-level approach, which consists of resampling techniques that are used to balance the data by either *undersampling* or *oversampling* the dataset [25].

First, *undersampling* is the process of decreasing the number of instances (or samples) in the majority classes. Some of the most commonly used undersampling methods consist of using the KNN algorithm, clustering or ensemble techniques. In the case of the KNN (k-nearest neighbors) algorithm, it is used to eliminate data where the target class is not equal to the majority of its "nearest neighbor instances" [25]. The use of the *k-means* clustering method aims at balancing the instances of imbalanced classes by reducing the number of majority instances [22]. In turn, in random *undersampling* methods [7], instances of majority classes are generally randomly sampled without label replacement to create a fully balanced training set [23]. Finally, there are assembly methods such as *EasyEnsemble* [31] where the majority class is divided into several subsets where the size of each subset is equal to the size of a minority class.

Secondly, *oversampling* consists of increasing the number of instances or samples of minority classes by producing new instances or repeating pre-existing ones. The most common technique is known as SMOTE (Synthetic Minority Over-sampling Technique) [8], where, to oversample, a sample is taken from the data set and the k nearest neighbors are considered based on the feature space, creating a synthetic data point from the multiplication of one of the feature vectors and a random value, usually between 0 and 1. Another example of oversampling methods is Borderline-SMOTE [17] whose objective is to identify minority samples located at the decision boundary and use them for oversampling, avoiding the potential risks of overgeneralization that occur with SMOTE. RAMO-Boost (Ranked Minority Oversampling in Boosting) [9] is a technique that systematically generates synthetic samples using an ordered sampling probability distribution. There are also other synthetic sample generation approaches, such as ADASYN [19] and MWMOTE [3], which have obtained good results based on modifications to the synthetic data generation mechanisms.

Finally, some studies have shown that the combination of oversampling and undersampling methods allows better classifier performance than methods used separately [8]. In any case, the number of approaches proposed to solve these problems allow us to infer the importance of the topic for the evolution of supervised machine learning techniques.

It is important to observe that techniques based on undersampling are not an alternative when minority classes have very few identified instances because it has been shown that to accurately characterize the effectiveness of such systems, they must be evaluated at the operational scale at which they will be used in practice [20]. On the other hand, most oversampling techniques are based on the generation of new synthetic instances that are not part of the real observations, which clearly seems to be a limitation.

However, fundamentally as a result of the massification of Internet access, millions and millions of data are generated every day, and the amount of data available for training classification algorithms is not a restriction [13]. The limitations here are given by the capacity to label those available data with the traditional (manual) strategy performed by a human. Hence, while expert labels provide the traditional cornerstone for training and evaluating classifier models, limited or expensive access to experts represents a practical bottleneck [20].

In that context, we present a new alternative for learning from unbalanced data that generates new training samples, not artificially, but by identifying unlabeled instances in the original dataset. In this paper, we present a new approach, previously used as a semi-supervised labeling strategy [16], which consists of starting from a manually labeled dataset and, using feature selection strategies, extracting representative terms from each minority class to retrieve new instances from a repository of unlabeled data and thus balancing the dataset with non-synthetic examples.

3 Research Methodology

As discussed above, the general objective of this research is to present a new strategy for learning from imbalanced data and to evaluate its performance for automatic email classification in relation to oversampling and undersampling strategies widely used in the scientific community. Figure 1 shows the schematic diagram of the process developed.

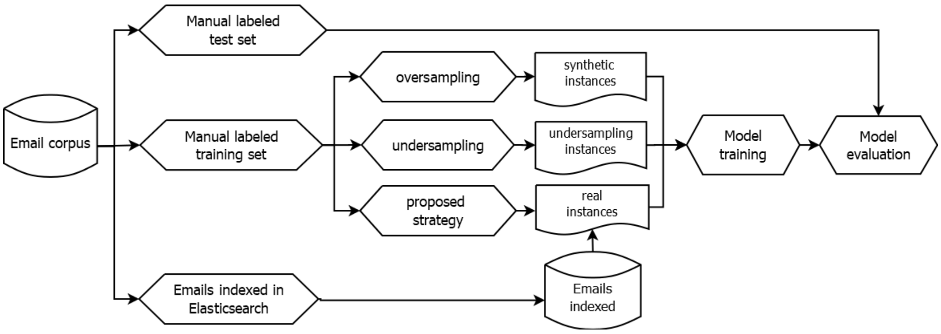

Fig. 1. Workflow proposed in this research

From an initial dataset, a subset of instances was selected and manually labeled by domain experts and two datasets were separated, one for model training and the other for evaluation. The oversampling and undersampling strategies were applied directly on the training set to consolidate the training data set.

For the new proposed strategy, both the manually labeled training set and the complete dataset were used, which was previously indexed in a general purpose search engine such as *Elasticsearch* for efficiency. In this case, the training set was further used to obtain the key terms representing each class of the problem and then retrieve unlabeled documents containing those terms to enrich the minority classes.

Once the datasets were consolidated with the class balancing strategies applied, classification models were trained and evaluated from the set reserved for this purpose. In the following sections, the most important issues related to the developed process are explained in greater detail.

3.1 Description of the dataset

For the experiments, were used a collection of 24700 e-mails generated from academic questions made by students to the administrative staff of the National University of Lujan. These questions are about procedures derived from the academic activity and the original e-mails were used without fixing any kind of typos or syntax errors. From those 24700 e-mails, 1000 were randomly selected and labeled around the question topic by a domain expert.

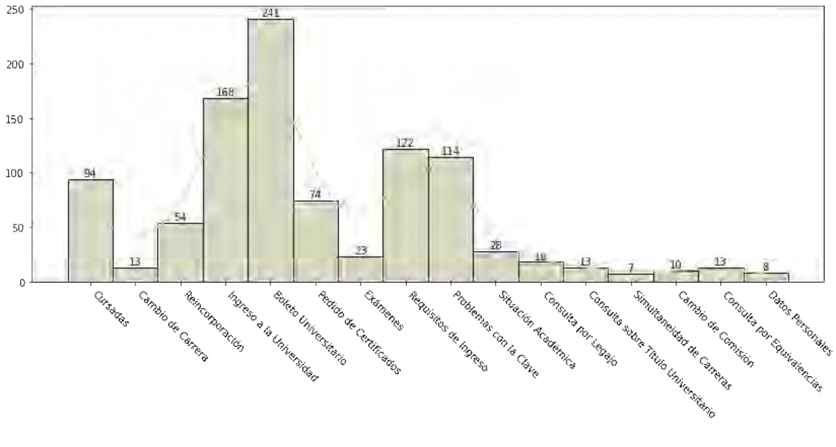

Fig. 2. Observed frequency for classes resulting from manual labeling

The 16 classes, all independent of each other, resulting from the labeling are: *Boleto Universitario, Cambio de Carrera, Cambio de Comisión, Consulta por Equivalencias, Consulta por Legajo, Consulta sobre Título Universitario, Cursadas, Datos Personales, Exámenes, Ingreso a la Universidad, Pedido de Certificados, Problemas con la Clave, Reincorporación, Requisitos de Ingreso, Simultaneidad de Carreras y Situación Académica.* The frequency distribution for each class is shown in Figure 2.

As it can be seen, the classes are highly imbalanced, an aspect that usually difficulties the classification process and that will be addressed with the process proposed in this paper.

3.2 Strategies used for the treatment of imbalanced datasets

The three strategies used for the treatment of the imbalanced data prior to the training of the automatic classification models are briefly presented below.

***Oversampling* Strategies.** The strategies implemented[3] were *RandomOverSampler*, *SMOTE*, *ADASYN* and *BorderSMOTE*. The first strategy, also known as ROSE (from the acronym for *random over sampling examples*), consists of generating new samples by random sampling with replacement of the current available samples and relies on a theoretical basis supported by the properties of kernel methods [24]. For its part, SMOTE (*Synthetic Minority Over-sampling Technique*) is one of the most recognized oversampling strategies, where, broadly speaking, the minority class is oversampled by introducing synthetic examples based on its k nearest neighbors depending on the amount of oversampling required [8]. In this sense, Borderline-SMOTE [17] is a variant of SMOTE that basically tries to determine the instances of the minority classes that are on the boundaries and generate synthetic instances from them. Finally, the essential idea of ADASYN (*Adaptive Synthetic Sampling*) [19] is to use a weighted distribution for the different examples of minority classes according to their level of learning difficulty, where more synthetic data is generated to examples of minority classes that are more difficult to learn.

***Undersampling* Strategies.** The strategies implemented were *RandomUnderSampler*, *ClusterCentroids* and *EditedNearestNeighbours*. The first strategy arbitrarily removes instances of the majority class in the training dataset [18] while in the case of the strategies based on *clustering* [22], a undersampling method based on replacing or removing instances by the centroids of the minority class instances is employed to reduce the number of majority class data samples.

On the other hand, the strategy *Edited Nearest Neighbours* [32] applies the nearest neighbor algorithm and "edits" the data set by removing samples that do not "sufficiently" match their neighborhood.

Proposed strategy. The strategy was initially presented as a semi-supervised classification strategy [16]. From an initial base with traditionally labeled mails, an extraction of the main features for each class is performed using different techniques, in this case TF-IDF and SS3 due to the results obtained in the previous work.

In the case of the TF-IDF technique [2], under this strategy, weighting per term grouped by class is used to determine which are the most important for each

[3] Implementations were performed with the **Imbalanced-learn** library for Python.

class. In the case of SS3 [4], it generates a function *gv(w, c)* that weights words relative to categories; to be more specific, *gv* takes a word w and a category c and generates a number in the interval [0,1] that represents the degree of confidence with which w belongs exclusively to c.

After retrieving the representative terms per class with both strategies, with the complete knowledge base indexed in a general purpose search engine such as *Elasticsearch*, documents from each class are retrieved based on the features detected by each technique and a new dataset is consolidated based on the instances that were retrieved by both feature selection strategies.

These instances are complemented by training dataset instances prior to the training of the classification model in order to balance it.

3.3 Generation of the Classification Models

As for classification techniques, support vector machines (SVM) were used because of their high performance for vectorized data, since vectorized data is generally required for the resampling strategies to be implemented.

SVM is a classical approach that has gained popularity over time due to some attractive features and its empirical performance. The main objective of support vector machines is to select the hyperplane which separates the training instances with a maximum distance criterion [27].

To evaluate the models, the remaining 200 manually labeled instances were reserved. Finally, the analysis of the selection of the generated models was performed based on the standard metrics *accuracy*, *precision* and *f1-score*.

4 Experiments

For the experiments[4], the training set with the 800 instances was used in all cases. Prior to training, queries were vectorized using 3-4 character grams and a TF-IDF weighting in all cases, and then class balancing strategies were applied. SVM was used in combination with a grid search alternating C (0.01, 0.1, 1), gamma (0.1, 0.01) parameters as well as kernels (rbf, linear, sigmoid), with and without class weighting.

It is important to clarify that in the case of the proposed strategy, 200 instances were retrieved for each class and feature selection technique from the database indexed in Elasticsearch, which resulted in a limitation because the number of instances resulting from the cross-linking between the instances retrieved by the two techniques meant that in some classes the amount of balance required for balancing was not reached, although the existing imbalance was reduced. This option was chosen over that of recovering a larger number of instances, with a lower coincidence *score*, in order not to introduce noise in the training set. To mitigate this situation, a variant of the proposed strategy is the definition of a smaller alternative N of instances, such as the average available per class, in order to reduce the distortion.

[4] Experiments available at `github.com/jumafernandez/imbalanced_data`

Next, classifiers were trained from the balanced datasets from the different strategies and the performance of the models was evaluated with the 200 instances reserved for this purpose, applying the accuracy, F1-score and precision metrics. The results are presented in Table 1.

Table 1. Experiments with class balancing techniques

Strategy	Accuracy	F1-Score	Precision
SVM (without class balancing)	0.810	0.80	0.82
RamdomOverSampler	0.810	0.80	0.81
SMOTE	0.805	0.79	0.81
ADASYN	0.810	0.80	0.81
BorderSMOTE	0.805	0.79	0.81
RamdomUnderSampler	0.660	0.68	0.73
ClusterCentroids	0.645	0.68	0.75
EditedNearestNeighbours	0.665	0.60	0.61
Proposed strategy	**0.820**	**0.83**	**0.85**
Proposed strategy (n = average = 115)	**0.820**	**0.83**	**0.85**

Based on the above experiments, it can be stated that none of the preexisting techniques, either oversampling or undersampling, were able to improve the results obtained with the original dataset with the highly imbalanced classes. On the other hand, it is observed that the proposed strategy improved all the metrics in both variants equally.

In turn, another advantage of the proposed strategy, by incorporating nonsynthetic instances to the training dataset, lies in the possibility of using it for new classification approaches based on neural networks, either those of deep learning as well as those based on transformers, a limitation that is observed in balancing strategies based on synthetic examples in general. The results of running the experiments in BERT (Bidirectional Encoder Representations from Transformers)[5] are transcribed below [12].

Table 2. Experiments with class balancing techniques with BERT

Strategy	Accuracy	F1-Score	Precision
BERT (without class balancing)	0.860	0.847	0.845
Proposed strategy	**0.865**	**0.865**	**0.878**
Proposed strategy (n = average = 115)	0.840	0.837	0.854

Table 2 shows that the proposed strategy is still effective but only for the conventional approach. In the case of the variant by the mean number of in-

[5] For model training, we experimented with a pre-trained model native to the Spanish language [6] and a set of hyperparameters successfully used in a previous work [15].

stances per class, the results are lower for the *accuracy* and *f1-score* metrics, between 1% and 2%, and higher in similar proportions for the *precision*.

5 Conclusions

This paper presents a novel strategy for learning from imbalanced data sets based on class oversampling by retrieving new unlabeled instances from a data repository of the same nature as the labeled data.

The fact that the instances for resampling come from real instances is presented as an advantage over strategies that generate synthetic samples. In principle, it may appear as a weakness to require an additional repository of data for experimentation. However, in full-scale problems of the real world it is normal to have a large -though unlabeled- data repository available.

Another advantage of the proposed strategy, by incorporating non-synthetic instances to the training data set, lies in the possibility of using it for new classification approaches based on neural networks, whether *deep learning* or *transformer-based*, a limitation that is observed in strategies based on synthetic examples in general.

Based on the results obtained, it can be concluded that this new strategy is competitive with respect to other resampling strategies widely used in the scientific community, either for traditional classification approaches, such as the one proposed for SVM, or for new transformer-based approaches, such as BERT.

Finally, although the present study has limited the experimentation to the domain of e-mail classification, we believe that the proposed strategy is generalizable to other domains where unlabeled text documents are available and, as future work, we propose to carry out further work applied to a more general text classification context.

References

1. Alghoul, A., Al Ajrami, S., Al Jarousha, G., Harb, G., Abu-Naser, S.S.: Email classification using artificial neural network. ACM (2018)
2. Bafna, P., Pramod, D., Vaidya, A.: Document clustering: Tf-idf approach. In: 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT). pp. 61–66. IEEE (2016)
3. Barua, S., Islam, M.M., Yao, X., Murase, K.: Mwmote–majority weighted minority oversampling technique for imbalanced data set learning. IEEE Transactions on knowledge and data engineering 26(2), 405–425 (2012)
4. Burdisso, S.G., Errecalde, M., Montes-y Gómez, M.: A text classification framework for simple and effective early depression detection over social media streams. Expert Systems with Applications 133, 182–197 (2019)
5. Cardenas, M.E., Castillo, J.J., Navarro, M., Hernández, N., Velazco, M.: Herramientas para el desarrollo de sistemas de análisis de textos no estructurados. In: XXI Workshop de Investigadores en Ciencias de la Computación (WICC 2019, Universidad Nacional de San Juan). (2019)

6. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
7. Chawla, N.V.: Data mining for imbalanced datasets: An overview. Data mining and knowledge discovery handbook pp. 875–886 (2009)
8. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357 (2002)
9. Chen, S., He, H., Garcia, E.A.: Ramoboost: Ranked minority oversampling in boosting. IEEE Transactions on Neural Networks 21(10), 1624–1642 (2010)
10. Chen, Z., Tao, R., Wu, X., Wei, Z., Luo, X.: Active learning for spam email classification. In: Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence. pp. 457–461 (2019)
11. Dalal, M.K., Zaveri, M.A.: Automatic text classification: a technical review. International Journal of Computer Applications 28(2), 37–40 (2011)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Fanny, F., Muliono, Y., Tanzil, F.: A comparison of text classification methods k-nn, naive bayes, and support vector machine for news classification. Jurnal Informatika: Jurnal Pengembangan IT 3(2), 157–160 (2018)
14. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from imbalanced data sets, vol. 10. Springer (2018)
15. Fernandez, J.M., Cavasin, N., Errecalde, M.: Classic and recent (neural) approaches to automatic text classification: a comparative study with e-mails in the spanish language. In: Short Papers of the 9th Conference on Cloud Computing, Big Data & Emerging Topics. p. 20 (2021)
16. Fernández, J.M., Errecalde, M.: Multi-class e-mail classification with a semi-supervised approach based on automatic feature selection and information retrieval. In: Conference on Cloud Computing, Big Data & Emerging Topics. pp. 75–90. Springer (2022)
17. Han, H., Wang, W.Y., Mao, B.H.: Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: International conference on intelligent computing. pp. 878–887. Springer (2005)
18. Hanafy, M., Ming, R.: Improving imbalanced data classification in auto insurance by the data level approaches. International Journal of Advanced Computer Science and Applications 12(6) (2021)
19. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). pp. 1322–1328. IEEE (2008)
20. Jung, H.J., Lease, M.: Evaluating classifiers without expert labels. arXiv preprint arXiv:1212.0960 (2012)
21. Lematre, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research 18(1), 559–563 (2017)
22. Lin, W.C., Tsai, C.F., Hu, Y.H., Jhang, J.S.: Clustering-based undersampling in class-imbalanced data. Information Sciences 409, 17–26 (2017)
23. Liu, B., Tsoumakas, G.: Dealing with class imbalance in classifier chains via random undersampling. Knowledge-Based Systems 192, 105292 (2020)
24. Menardi, G., Torelli, N.: Training and assessing classification rules with imbalanced data. Data mining and knowledge discovery 28(1), 92–122 (2014)

25. Mohammed, R., Rawashdeh, J., Abdullah, M.: 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. IEEE (2020)
26. Shelke, M.S., Deshmukh, P.R., Shandilya, V.K.: A review on imbalanced data handling using undersampling and oversampling technique. Int. J. Recent Trends Eng. Res 3(4), 444–449 (2017)
27. Skiena, S.S.: The data science design manual. Springer (2017)
28. Tang, G., Pei, J., Luk, W.S.: Email mining: tasks, common techniques, and tools. Knowledge and Information Systems 41(1), 1–31 (2014)
29. Usai, A., Pironti, M., Mital, M., Mejri, C.A.: Knowledge discovery out of text data: a systematic review via text mining. Journal of knowledge management (2018)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
31. Wang, T., Lu, C., Ju, W., Liu, C.: Imbalanced heartbeat classification using easyensemble technique and global heartbeat information. Biomedical Signal Processing and Control 71, 103105 (2022)
32. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics (3), 408–421 (1972)