Knowledge and Information Systems (2023) 65:31–57
https://doi.org/10.1007/s10115-022-01772-8

SURVEY PAPER

Imbalanced data preprocessing techniques for machine learning: a systematic mapping study

Vitor Werner de Vargas[1] · Jorge Arthur Schneider Aranda[1] · Ricardo dos Santos Costa[2] · Paulo Ricardo da Silva Pereira[2] · Jorge Luis Victória Barbosa[1,2]

Received: 8 October 2021 / Revised: 27 September 2022 / Accepted: 2 October 2022 /
Published online: 9 November 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract
Machine Learning (ML) algorithms have been increasingly replacing people in several application domains—in which the majority suffer from data imbalance. In order to solve this problem, published studies implement data preprocessing techniques, cost-sensitive and ensemble learning. These solutions reduce the naturally occurring bias towards the majority sample through ML. This study uses a systematic mapping methodology to assess 9927 papers related to sampling techniques for ML in imbalanced data applications from 7 digital libraries. A filtering process selected 35 representative papers from various domains, such as health, finance, and engineering. As a result of a thorough quantitative analysis of these papers, this study proposes two taxonomies—illustrating sampling techniques and ML models. The results indicate that oversampling and classical ML are the most common preprocessing techniques and models, respectively. However, solutions with neural networks and ensemble ML models have the best performance—with potentially better results through hybrid sampling techniques. Finally, none of the 35 works apply simulation-based synthetic oversampling, indicating a path for future preprocessing solutions.

✉ Vitor Werner de Vargas
vitorwv@edu.unisinos.br

Jorge Arthur Schneider Aranda
jsaranda@unisinos.br

Ricardo dos Santos Costa
ricsantos@edu.unisinos.br

Paulo Ricardo da Silva Pereira
prpereira@unisinos.br

Jorge Luis Victória Barbosa
jbarbosa@unisinos.br

1 Applied Computing Graduate Program, University of Vale do Rio dos Sinos, São Leopoldo, Rio Grande do Sul 93022-750, Brazil

2 Electrical Engineering Graduate Program, University of Vale do Rio dos Sinos, São Leopoldo, Rio Grande do Sul 93022-750, Brazil

Keywords Imbalanced data · Preprocessing techniques · Sampling · Machine learning · Systematic mapping study

1 Introduction

Machine Learning (ML) has been increasingly applied to domain areas in which data is available for process automation. However, the training process is challenging since ML algorithms conceptually learn from balanced distributions [1]. Therefore, learning from unevenly distributed samples can decrease both accuracy and reliability from the trained model. This characteristic is called imbalance or unbalance [2].

Imbalanced data occur naturally in the majority of real-world problems. Nevertheless, when the ratio between the minority and majority—Imbalance Ratio (IR)—is low, the minority class tends to be ignored as noise [3]. Consequently, the ML model becomes biased towards the majority class, leading to more False Positives (FP) and less True Positives (TP) [4].

The solution for imbalanced data applications can be implemented in two levels [5]:

- Data: preprocessing data before learning through algorithms for undersampling the majority sample, oversampling the minority sample, or both (hybrid sampling)—as illustrated in Fig. 1;
- Algorithmic: processing learning through algorithms optimized for imbalanced data, such as cost-sensitive and ensemble ML models.

Algorithmic approaches optimize learning for specific application characteristics, being hard to reapply models to other datasets. Conversely, data level solutions fix the imbalance and allow the use of standard ML models [6]. Additionally, data level solutions enable implementations in conjunction with ensemble ML models—further improving learning [7].

This study's main objective is to review papers solving ML in imbalanced data applications through data level preprocessing techniques. Additionally, this paper details the analyzed works' domain areas and solutions—specifying current and effective sampling techniques and ML models, and checking the use of simulation data, thus serving as a basis for future works. Structured as a systematic mapping study, the search process found 9927 papers through 7 digital libraries. From these, an eight-step filtering process selected 35 papers for analyses and discussions.

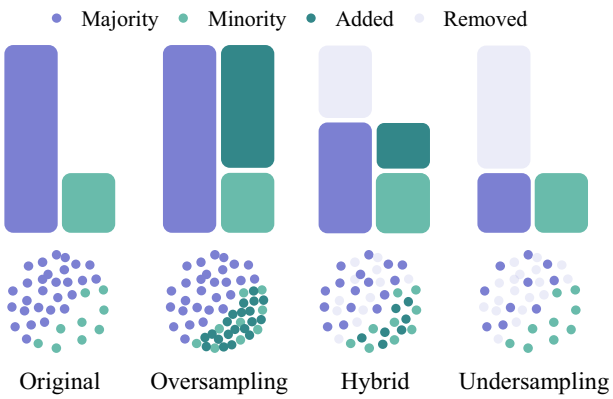

Fig. 1 Sampling types for imbalanced data preprocessing

The paper is organized as follows: Sect. 2 describes related works and this study's contribution; Sect. 3 details the materials and methods used in this literature review; Sect. 4 answers research questions, discusses results, and presents taxonomies and illustrations of the findings; and, finally, Sect. 5 provides conclusions and lessons learned from the study.

2 Related works

The research method described in Sect. 3 yielded 5 reviews and surveys addressing techniques for dealing with the imbalance problem generally [8–12]. Additionally, 19 reviews analyzed solutions limited to specific applications [13–31].

This section describes general and application-limited reviews in Sects. 2.1 and 2.2, respectively. Moreover, Sect. 2.3 details this study's contribution.

2.1 General reviews and surveys

Kaur et al. [8] presented an in-depth literature review on the imbalanced data challenges for ML. The paper extensively details solution methods in ML, exploring preprocessing techniques, cost-sensitive learning, algorithm-centered and hybrid methods. The authors structured and analyzed works through domain areas and corresponding applications. Additionally, the authors described and compared ML algorithms applied to metrics obtained in the selected studies.

Felix and Lee [9] reviewed published studies on preprocessing techniques for general ML applications. The work focuses on evaluating the quality of published papers, highlighting the score per data-related issues and preprocessing techniques—hence directing future works.

Spelmen and Porkodi [10] detailed solutions from papers handling imbalanced data on both data and algorithmic levels—including hybrid models. The study describes the proposed solution and results for each work through a discussion organized by solution methods.

Susan and Kumar [11] surveyed studies on preprocessing techniques for ML applications. The paper thoroughly describes sampling methods and how each analyzed work implemented the proposed solutions. Finally, the survey also summarizes experimental procedures, details, and reported results.

Shakeel et al. [12] reviewed works on preprocessing techniques for ML binary and multiclass classification. The authors briefly described classification algorithms, preprocessing, and ensemble methods.

Furthermore, the reviewed papers [8–12] discuss strengths, weaknesses, applications, and opportunities for future works. Table 1 outlines relevant topics of these papers: publication year, data level preprocessing as the only solution method, ML-only applications, Quality Assessment (QA), and primary focus. The topic is classified as "partially" when the study covers other balancing solutions, such as cost-sensitive and ensemble learning, or applications without ML.

2.2 Application-focused reviews and surveys

The research method also found 19 reviews addressing solutions for specific imbalanced data and ML applications. These papers explore: classification algorithms [13–15], credit risk evaluation [16], disease diagnosis [17–23], fault diagnosis [24, 25], transaction fraud detection [26, 27], software defect prediction [28–30], and spam filtering [31]. Furthermore,

Table 1 Details of topics from related works

Work	Year	Preprocessing	ML	QA	Primary focus
[8]	2019	Partially	Yes	No	Applications and results
[9]	2019	Yes	Partially	Yes	Quality Assessment
[10]	2018	Partially	Partially	No	Solution description
[11]	2020	Yes	Yes	No	Solution description and results
[12]	2017	Partially	Yes	No	Classification applications

some of these papers also limit reviewed studies by the ML algorithms, covering only boosting [14], Convolutional Neural Networks (CNN) [15], and Deep Learning (DL) [13, 21].

2.3 Contribution

Although there is a large number of studies addressing imbalanced data through preprocessing, Felix and Lee [9] affirm that there is a lack of literature reviews in order to assert the reliability of the proposed techniques.

Related works mainly focus on specific applications [13–31]. Additionally, other works focus on describing the solutions proposed by the reviewed papers [8, 10–12], or assessing their quality [9].

Conversely, this paper aims to quantitatively detail sampling techniques and ML models in imbalanced data applications. This approach centers on structuring and analyzing publication data from different domains. In this sense, the study enables the creation of 2 taxonomies of sampling techniques and ML models tested in the reviewed studies. Additionally, this analysis may outline novel findings on performance and correlation with domain areas.

The quantitative analysis evaluates the reliability of sampling techniques and ML models through the number and relative performance by comparing the ratio between selected and tested methods in the reviewed studies. Moreover, this study searches for simulation-based solutions as support for future implementations.

Expanding related reviews from Table 1, this study covers both preprocessing and ML, assessing the studies' quality through answers for the Research Questions (RQs). Finally, the publication date gap may also contribute by including recent studies. Related works covered their most recent papers from 2017 [10, 12], 2018 [8, 9], and, more recently, 2020 [11].

3 Research method

This paper applied a systematic mapping methodology for conducting an evidence-based literature review of research publications addressing preprocessing techniques for imbalanced data in ML applications. Generally used to identify, aggregate, and classify studies on the research topic, the methodology aims to be unbiased and replicable [32, 33].

Oriented by the guidelines proposed by Petersen et al. [34], this systematic mapping defined the following procedures: (1) Research Questions; (2) Search strategy; (3) Papers filtering; and (4) Quality Assessment.

Table 2 Research questions

RQ#	Description
GQ1	How have preprocessing techniques been used to optimize Machine Learning from imbalanced datasets?
FQ1	What are the domain areas of Machine Learning applications with imbalanced datasets?
FQ2	Which preprocessing techniques are used to balance imbalanced datasets for Machine Learning training?
FQ3	Are there any studies that use simulation data for preprocessing imbalanced datasets?
FQ4	Which Machine Learning models are used in imbalanced data applications?
FQ5	Which development tools are used for implementing the proposed solutions? (programming language, package, or software)
FQ6	Are there any correlations between domain areas and preprocessing techniques or Machine Learning models?
SQ1	How has the quantity of studies evolved? (publications per year)
SQ2	Where have the studies been published? (type of venue and digital library)

3.1 Research questions

Accurate RQs are the key to finding a good sample of articles on a domain area [35]. Hence, a preliminary research and analyses of the resulting articles defined this study's questions. These questions guided the discovery and characterization of studies applying sampling techniques for improving ML applications with imbalanced datasets.

This study divided RQs into three sets, shown in Table 2:

- General Question (GQ): it states the main research focus;
- Focused Questions (FQs): these 6 questions detail existing solutions in order to structure models, identify patterns, limitations, and gaps for future research;
- Statistical Questions (SQs): these 2 questions comprise bibliography information for chronological analysis and QA.

3.2 Search strategy

The study defined three steps for the search strategy: (1) specify search string; (2) select databases; and (3) collect results. The first step identified the major terms and their most relevant synonyms—based on preliminary research and related works. Subsequently, the search string merged the major terms with their synonyms with Boolean operators. Table 3 presents the specified string.

The preliminary research found other combinations of search terms yielding numerous results—such as "filtering" for "preprocessing", and "class imbalance" for "imbalanced data". However, these synonyms created negative effects. For instance, "filtering" resulted

Table 3 Search string

Major term	Search terms
Imbalanced data	((“imbalanced data” OR “imbalanced dataset” OR “imbalanced data set” OR “unbalanced data” OR “unbalanced dataset” OR “unbalanced data set”)
	AND
Preprocessing	(preprocessing OR pre-processing OR preparation)
	AND
Machine Learning	(“machine learning” OR “deep learning” OR “artificial intelligence”))

in too many irrelevant signal noise reduction works, and “class imbalance” biased results towards general classification problems.

Secondly, the search strategy encompassed 7 digital libraries: Association for Computing Machinery (ACM), IEEE Xplore, Institution of Engineering and Technology (IET), Science Direct, Scopus, Springer Link, and Wiley. The selection of these libraries prioritized well-known research sources with multidisciplinary fields, which is essential to finding applications in various areas of knowledge—as suggested by Silva and Braga [36].

Finally, in addition to the search string as the search query, the research applied filters for language and type of venue according to the filtering process—when available in the digital library.

3.3 Papers filtering

The collected papers went through a filtering process, removing studies unrelated to sampling techniques for ML applications. The following Exclusion Criteria (EC) supported the filtering process:

- EC1: The study is not written in English;
- EC2: The study venue is neither conference nor journal;
- EC3: The study matches the keywords defined in the search string, but the context is different from the research purposes;
- EC4: The study is a literature review (Sect. 2);
- EC5: The study is not accessible in full-text;
- EC6: The study is a short paper (4 pages or less);
- EC7: The solution focuses on algorithmic level techniques for imbalanced data;
- EC8: The study does not detail the sampling techniques or ML models implemented in the solution—answering FQ2 and FQ4;
- EC9: The study validates the proposed solution through datasets from multiple applications.

Papers filtering started at the initial search from each digital library, removing results complying with EC1 and EC2. This process did not have any date restraint, therefore collecting all results published in conferences or journals, and written in English. Then, one filter by title and one filter by abstract extracted studies meeting EC3 and EC4. After that, a combination of the remaining papers removed repeated works.

Table 4 Quality scores for the answers of research questions

Answer	Score	Criterion
Y	1.0	The paper entirely answers the question
P	0.5	The paper partially answers the question
N	0.0	The paper does not address the topic

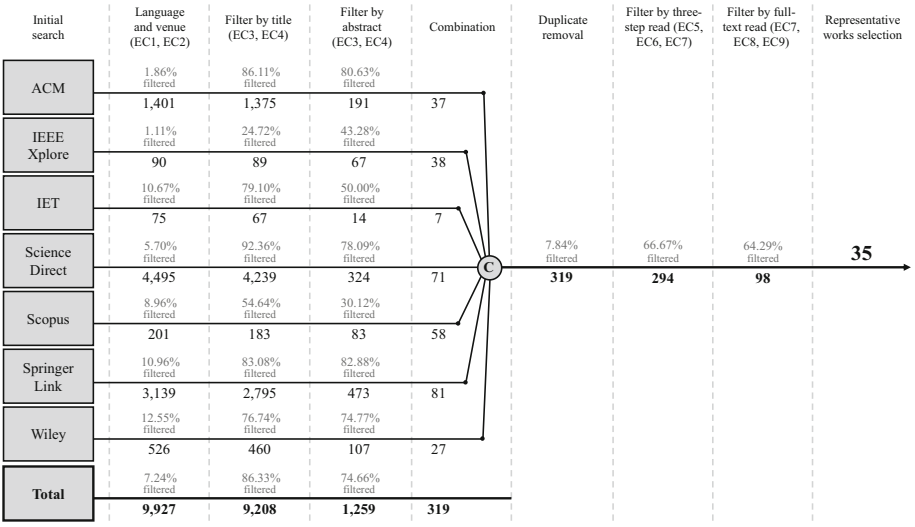

Fig. 2 Filtering process

Subsequently, the original papers went through a filter based on the three-pass method [37], excluding papers complying with EC5, EC6, and EC7. Finally, a careful full-text read selected the most representative works for the research purposes. The final step rejected algorithmic level solutions, low-quality papers, and papers without a single predetermined application—meeting EC7, EC8, or EC9.

3.4 Quality Assessment

Following the scoring system proposed by Kitchenham et al. [32], this paper evaluates the selected papers' quality applying FQs 1 to 5—since they inherently structure the research. Table 4 presents the QA scores, attributing better values for more satisfactory answers through a classification between Yes (Y), Partially (P), and No (N). Additionally, this study also presents the H-Index, year of publication, and type of venue of each paper.

4 Results

The collection of results in all 7 digital libraries integrated 9927 studies. After an eight-step filtering process, the selection of the representative works resulted in the 35 papers indicated in the QA (Table 5). Figure 2 details the filtering process.

Table 5 Quality assessment

ID	Ref.	Year	Venue	FQ1–FQ4	FQ5	QA	H-Index
1	[68]	2020	Journal	1.0	0.0	4.0	110
2	[65]	2020	Journal	1.0	0.0	4.0	127
3	[57]	2019	Journal	1.0	0.0	4.0	119
4	[47]	2020	Conference	1.0	0.0	4.0	20
5	[48]	2019	Conference	1.0	0.0	4.0	–
6	[39]	2020	Journal	1.0	1.0	5.0	68
7	[56]	2018	Journal	1.0	0.0	4.0	38
8	[52]	2016	Journal	1.0	1.0	5.0	180
9	[63]	2019	Conference	1.0	1.0	5.0	–
10	[46]	2016	Conference	1.0	0.0	4.0	–
11	[53]	2017	Conference	1.0	0.0	4.0	–
12	[42]	2019	Conference	1.0	0.0	4.0	–
13	[54]	2020	Conference	1.0	0.0	4.0	–
14	[64]	2020	Conference	1.0	0.5	4.5	–
15	[55]	2019	Journal	1.0	0.0	4.0	71
16	[41]	2019	Journal	1.0	0.0	4.0	22
17	[62]	2015	Journal	1.0	0.5	4.5	103
18	[60]	2020	Journal	1.0	0.5	4.5	66
19	[40]	2021	Journal	1.0	0.5	4.5	70
20	[50]	2020	Journal	1.0	1.0	5.0	43
21	[59]	2020	Journal	1.0	0.0	4.0	18
22	[43]	2019	Journal	1.0	0.5	4.5	44
23	[44]	2013	Journal	1.0	0.0	4.0	108
24	[51]	2019	Journal	1.0	0.0	4.0	52
25	[66]	2019	Conference	1.0	1.0	5.0	–
26	[58]	2021	Journal	1.0	1.0	5.0	130
27	[49]	2019	Journal	1.0	1.0	5.0	143
28	[38]	2006	Journal	1.0	0.0	4.0	87
29	[71]	2019	Conference	1.0	0.5	4.5	76
30	[70]	2017	Journal	1.0	0.5	4.5	92
31	[67]	2020	Journal	1.0	0.0	4.0	184
32	[61]	2019	Journal	1.0	0.0	4.0	4
33	[45]	2020	Journal	1.0	0.0	4.0	102
34	[72]	2013	Journal	1.0	1.0	5.0	121
35	[69]	2020	Journal	1.0	0.0	4.0	87

The selected studies completely answer FQs 1 to 4, so Table 5 merges their quality score in the column "FQ1-FQ4". The only majorly unanswered question details development tools (FQ5). Therefore, all works have their QA between 4 and 5. This result indicates good quality papers—detailing application, sampling techniques, and ML models.

4.1 GQ1: How have preprocessing techniques been used to optimize Machine Learning from imbalanced datasets?

Data preparation is fundamental for ML. Hence, several preprocessing techniques can be applied to improve the learning process in applications with imbalanced datasets.

Cohen et al. [38] published the first study filtered in the search process. The authors proposed the use of two clustering techniques in a hybrid model: Agglomerative Hierarchical Clustering (AHC)-based oversampling and K-Means-based undersampling. Tested against Random UnderSampling (RUS) and Random OverSampling (ROS), the hybrid model achieved the most effective results with 5 different ML models—improving hospital-acquired (nosocomial) infection prediction.

Lee and Kim [39] also compared RUS, ROS, and a hybrid approach (RUS+ROS) with different sampling probabilities for DL-based toxicity classification in nuclear receptor compounds. The hybrid model enhanced specificity and sensitivity without compromising accuracy for two models—SCFP and FP2VEC.

Other works also create hybrid models combining RUS and oversampling through synthetic sample generation. Mahadevan and Arock [40] advanced ensemble learning by using RUS and Synthetic Minority Oversampling TEchnique (SMOTE). The system achieved the best results for review rating prediction in e-commerce – compared to other models. RUS+SMOTE avoided induced bias and loss of useful information.

Complementary hybrid models applied clustering techniques for undersampling with synthetic oversampling. Rustam et al. [41] applied Edited Nearest Neighbour (ENN) and SMOTE for improving the performance of cerebral infarction detection in hospital patients through SVM. The experimental results show that the performance of Support Vector Machine (SVM) classifiers is improved by using these techniques—which produce better accuracy as a hybrid algorithm rather than individually.

Similarly, Chang et al. [42] implemented hybrid sampling with ENN and ADAptive SYNthetic sampling (ADASYN) for enhancing the sensitivity of fraud identification in telephones through Stacked-SVM. Han et al. [43] developed a credit scoring solution preprocessed by a Gaussian Mixture Model (GMM)-based majority undersampling and SMOTE. Based on tested ML metrics with both Logistic Regression (LR) and Decision Trees (DT), the authors assessed that the proposed algorithm generally performs better than 11 standard sampling algorithms.

Marqués et al. [44] also proposed credit scoring solutions by testing 8 undersampling and oversampling techniques with LR and SVM. The authors concluded that oversampling generally outperforms undersampling for both ML models. Following a congruent path, Pereira et al. [45] compared 8 well-known sampling techniques in order to identify COVID-19 from a record of chest X-Ray images. The most effective combination results from ENN with a Multi-Layer Perceptron (MLP) model.

Vu et al. [46] tested different techniques for encrypted network traffic identification. The study shows that ConDensed Nearest Neighbour (CDNN) and SVM-based SMOTE (SVM-SMOTE) performed the best as undersampling and oversampling techniques, respectively. However, both techniques proved to be slow compared to simpler algorithms, such as RUS, ROS, and SMOTE. Correspondingly, Shamsudin et al. [47] also achieved one of the highest precision and recall with a hybrid model between SVM-SMOTE and RUS—for credit card fraud detection with Random Forest (RF).

Haldar et al. [48] addressed epilepsy detection by applying the hybrid sampling technique Selective Preprocessing of Imbalanced Data, also known as SPIDER, with 3 different ML

models. The results showed that SPIDER with the K-Nearest Neighbours (KNN) classifier achieved the best performance.

Malhotra published two studies on software source code problems [49, 50]. The first, with Kamal [49], implements a modified version of the SPIDER2 algorithm, called SPIDER3. The proposed solution for software defect prediction performed better than SPIDER2 and the original SPIDER. However, ADASYN achieved the best average results in combination with 5 ML models.

In addition, Malhotra and Lata [50] performed an empirical study for selecting the best well-known sampling techniques and ML models for software maintainability prediction. After conducting tests with 14 techniques and 8 models, the authors found that Safe Level SMOTE (SL-SMOTE) significantly outperformed other techniques. The study also achieved relevant results with hybrid sampling between ENN and SMOTE, as well as Tomek Links (TL) and SMOTE.

Ma et al. [51] improved SL-SMOTE through an evolutionary optimization process for the algorithm's parametrization. The solution, named Evolutionary SL-SMOTE (ESL-SMOTE), achieved the highest metrics for seminal quality prediction with AdaBoost against related works. Additionally, the results indicate that the preprocessing technique achieves good recall for other models—such as Back Propagation Neural Networks (BPNN) and SVM.

Five works applied only SMOTE for improving ML and retaining superior overall results [52–56]. Yan et al. [52] achieved good results for lung cancer recurrence prediction with Gaussian Radial Basis Function Network (GRBFN). Moreover, Purnami and Trapsilasiwi [53] advanced breast cancer malignancy classification from biopsy records through Least Squares SVM (LS-SVM).

Another two SMOTE-focused studies used SMOTE in biology applications. Dewi et al. [54] improved stability of patchouli (flowering plants) classification with Extreme Learning Machine (ELM). Additionally, Zhang et al. [55] achieved higher accuracy for Protein-Protein Interactions (PPI) hot spots identification than related works through SMOTE and RF.

Gicić and Subasi [56] applied SMOTE in order to improve credit scoring for micro-enterprises of the minority class (poor). After preprocessing at 100% and 200% of the minority sample and testing with 15 classical and ensemble ML models, the authors concluded that the minority classification improved significantly and retained superior results overall.

Tra et al. [57] introduced a solution for diagnosing fault symptoms in the insulation oil of power transformers. The authors implemented an algorithm for improving SMOTE by estimating a local reachability distance of the majority and minority samples with two clusters. The Adaptive SMOTE (ASMOTE) algorithm achieved a higher classification accuracy than ROS and SMOTE with the proposed MLP model.

Comparably, Jiang and Li [58] improved fault detection in wind turbines by combining Dependent Wild Boostrap (DWB) with SMOTE (DWB-SMOTE). Since wind buffers have multivariate time-series of sensors from several subsystems, the proposed CNN model generated better temporal-dependent synthetic samples and, consequently, better results.

Faris et al. [59] tested various oversampling techniques and ML models in order to predict companies' financial bankruptcy through financial and non-financial records. After analyzing the results, the authors concluded that SMOTE with AdaBoost achieved promising and reliable predictions.

A modified version of SMOTE, called BorderLine SMOTE (BL-SMOTE), focuses on synthetic sample generation at the boundary between classes. Smiti and Soui [60] proposed this technique for companies' financial bankruptcy prediction through DL. Jiang et al. [61] also applied BL-SMOTE for heartbeat classification through electrocardiograms with CNN. Both works achieved the best results with BL-SMOTE.

Santos et al. [62] implemented a clustering-based oversampling approach through K-Means++ and SMOTE for hepatocellular carcinoma survival prediction. ML with Artificial Neural Networks (ANN) and LR presented significantly better results than without clustering or oversampling. Alternatively, Tashkandi and Wiese [63] applied K-means++ for undersampling. The results indicated an improvement in the prediction accuracy of mortality risk prediction in Intensive Care Units (ICUs) through different classical and ensemble ML models.

Zhou et al. [64] undersampled standard features for lower back pain early diagnosis through K-Means clustering—testing both stratified sampling and Manhattan distance. In general, these techniques improved the performance of all tested models for different "k" values.

Three papers proposed synthetic oversampling through a recent technique based on ML, called Generative Adversarial Networks (GAN) [65–67]. Liu et al. [65] developed GAN for balancing individual and fused sensor data of rotating machinery, such as bearing and gearbox. After learning with a multi-class CNN, the proposed techniques showed effective results in a wide range of IRs.

In addition, Gangwar and Ravi [66] applied GAN and Wasserstein GAN (WGAN) oversampling for a highly imbalanced dataset of credit card transactions. According to the authors, the results against ROS, SMOTE, and ADASYN indicate that GAN-based methods control FP spectacularly without affecting TP—which is essential for imbalanced data applications.

Yan et al. [67] implemented a Conditional WGAN (CWGAN) framework for multi-class air handling units' fault detection. Combined with quality control of the synthetic samples, the solution improved results from different ML classifiers—reaching an accuracy of almost 1 for every model.

Data spatial distribution is important for optimized classification. Therefore, Wang and Ye [68] implemented a spatial distribution-based sample generation for balancing historical and simulated power system stability data. The solution classifies distance intervals through KNN and creates properly distributed synthetic data through SMOTE—which feeds a Deep Neural Network (DNN) for evaluating transient stability.

Nnamoko and Korkontzelos [69] also created an optimized version of SMOTE for enhancing diabetes prediction. The algorithm uses the InterQuartile Range (IQR) technique for oversampling dispersed/extreme data before SMOTE, improving the training sample distribution. According to the authors, IQR+SMOTE consistently produced the best accuracy for different models and maintained the best overall metrics.

Liu et al. [70] introduced a Fuzzy-based OverSampling (FOS) algorithm for balancing tweets' data in spam detection—optimizing the distribution in synthetic sampling. The method improved precision for different ensemble learning models. However, ROS and RUS achieved better accuracy.

Filho et al. [71] studied automated essay scoring through ML regression and classification for Brazil's National High School Examination (ENEM). After testing SMOTE, ADASYN, ROS, and RUS, the authors concluded that random sampling performs better because the employed vectorization for feature extraction has unusual spatial characteristics.

Lastly, Zhou [72] tested different preprocessing techniques in order to enhance corporate bankruptcy prediction through ML. The authors concluded that there is no significant difference between the results of oversampling and undersampling with large amounts of data—for instance, in a dataset of USA companies from 1981 to 2009. However, the computational time is better in undersampling. When there is not much data, SMOTE performs the best overall. Additionally, GMM-based undersampling and RUS are better than Cluster Centroid (CC).

Table 6 Domain areas of the reviewed applications

Domain	Subdomain	Application	ID
Health (34.3%)	Cancer (8.6%)	Lung cancer recurrence	8
		Breast cancer malignancy	11
		Hepatocellular carcinoma survival	17
	Hospital (14.3%)	Risk of mortality in ICUs	9
		Cerebral infarction	16
		Nosocomial infections	28
		Heartbeats	32
		COVID-19	33
	Others (11.4%)	Epileptic seizure	5
		Lower back pain	14
		Seminal fluids quality	24
		Diabetes	35
Finance (22.9%)		Companies bankruptcy	18, 21, 34
		Credit cards fraud	4, 25
		Credit risk	7, 22, 23
Engineering (14.3%)	Fault (14.3%)	Power systems stability	1
		Rotating machinery	2
		Power transformers	3
		Wind turbines	26
		Air handling units	31
Biology (8.6%)		Nuclear receptor compounds toxicity	6
		Flowering plants species	13
		PPI hot spot	15
Software (8.6%)	Source code (5.7%)	Maintainability	20
		Defect	27
	Others (2.9%)	Network traffic data	10
Others (11.4%)		Telephone fraud	12
		E-commerce products rating	19
		Essay score	29
		Spam in tweets	30

4.2 FQ1: What are the domain areas of Machine Learning applications with imbalanced datasets?

There are 5 central domain areas for 31 of the reviewed works: health, finance, engineering, software, and biology. Additionally, 4 works are from other areas—classified as "others". Table 6 summarizes the domain areas and corresponding applications.

Health is the most prevalent domain, accounting for 12 studies. These studies differ in their application and type of classification. For instance, the 3 cancer-related works classify breast cancer malignancy [53], predict hepatocellular carcinoma survival [62], and predict lung cancer recurrence [52].

Nosocomial studies spread even more, proposing solutions for predicting risk of mortality in ICUs [63] and nosocomial infections [38], classifying heartbeats [61], and detecting cerebral infarction [41].

A recent work also detects COVID-19 from chest X-ray images [45]. Other health domain works introduce solutions such as detection of epileptic seizure [48] and lower back pain [64], as well as prediction of semen quality [51] and diabetes [69].

Finance, on the other hand, deals with cost-effective correlated problems. Representing 8 works, they predict companies bankruptcy [59, 60, 72], credit risk [43, 44, 56], and credit card fraud [47, 66].

Similarly, engineering studies propose solutions for fault diagnosis in different electrical and mechanical engineering applications. Accounting for 5 works, the solutions improve stability in power systems [68], wind turbines [58], power transformers [57], rotating machinery [65], and air handling units [67].

Furthermore, there are 3 papers related to software. These works improve ML for source code maintainability [50] and defect prediction [49], as well as network traffic data classification [46].

Biology also accounts for 3 studies. These studies introduce nuclear receptor compounds toxicity prediction [39], flowering plants species [54] and PPI hot spot classification [55].

Finally, 4 papers from other areas deal with telephone fraud detection [42], e-commerce rating prediction [40], essay score classification [71], and spam detection in tweets [70].

Dataset characteristics—such as features and IR—differ for each subdomain according to its applications. The reviewed studies do not always explore these characteristics, difficulting a comparative analysis. Specifically, some applications do not have enough data to infer the exact IR. Therefore, 4 works overcame this problem and generalized their solution by manually testing different IRs [39, 44, 65, 68].

Moreover, every domain area has particularities in its applications, demanding specialized preprocessing procedures before sampling. For instance: time-series data in engineering [58, 65, 67, 68] and health [61]; image processing in health [38, 45, 52] and biology [54]; text processing in other areas [40].

4.3 FQ2: Which preprocessing techniques are used to balance imbalanced datasets for machine learning training?

The literature covers a wide variety of preprocessing techniques for ML applications with specific characteristics and applications. This question focuses on sampling techniques for balancing datasets before ML training. Consequently, preprocessing techniques for other purposes, such as feature extraction, image, and natural language processing are not answered in this section.

Some of the reviewed studies propose a sampling technique and compare them with alternatives. Conversely, other reviews implement empirical analyses comparing several techniques to discuss results and select the best one(s). Therefore, this systematic mapping classified the techniques applied in each paper between "proposed", "compared" and "selected". Figure 3 shows a taxonomy of all sampling techniques, either proposed or compared in the reviewed papers—indicated by ID below the corresponding box. The taxonomy divides these algorithms into three types: "oversampling", "undersampling", and "hybrid sampling". Each algorithm is distributed according to its parent technique or type.

Additionally, Fig. 4a details the number of papers applying each technique in three columns: proposed, compared, and selected. The figure presents the most used techniques as

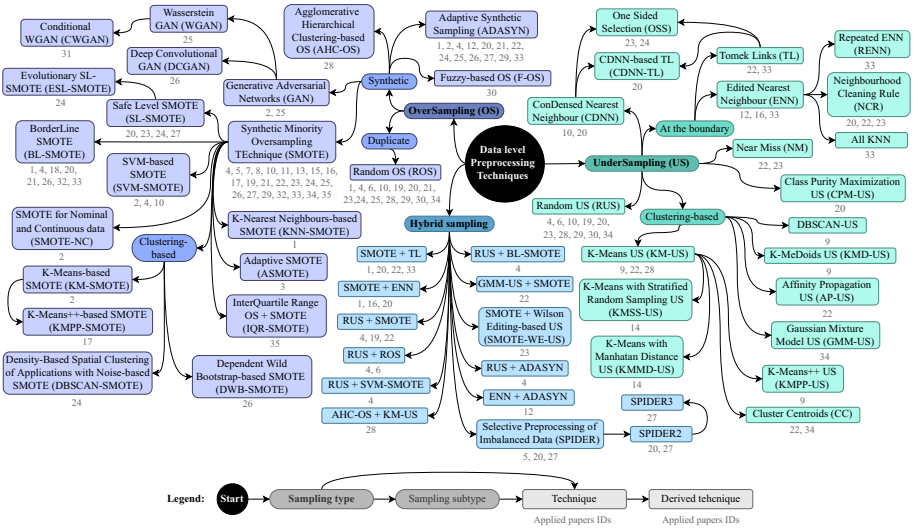

Fig. 3 Taxonomy of sampling techniques proposed or compared in the reviewed papers by ID

darker, while less used as lighter—based on a grayscale. These techniques are grouped by their types and subtypes, following the taxonomy in Fig. 3. The rightmost column indicates the percentage ratio of selected to proposed and compared techniques (relative performance).

The distribution of techniques in Fig. 4a shows a more significant interest in oversampling and hybrid sampling for the proposed solutions. The studies frequently compare results with standard oversampling and undersampling techniques—such as SMOTE, ADASYN, ROS, and RUS. Namely, each has at least 10 implementations.

Techniques focused at the boundary between classes are also popular (BL-SMOTE, ENN, and TL). Additionally, clustering-based algorithms are common in both oversampling and undersampling. For instance, AHC, KM, and DBSCAN have implementations in both. Nevertheless, clustering is more frequent in undersampling due to the grouping behavior.

In addition, the distribution of selected methods indicates growth in hybrid sampling and a decrease in oversampling and undersampling – relative to the proposed methods. Specifically, 13 of 35 papers tested hybrid sampling, out of which 9 (69.2%) were the best performing sampling type [38–44, 47, 48]. In the remaining 4 papers, hybrid sampling is outperformed by oversampling with KNN-SMOTE [68], SL-SMOTE [50], and ADASYN [49], and by undersampling with ENN [45].

However, oversampling remains the most selected sampling method, proportionally. Overall, synthetic sample generation techniques have the best performance, either individually or in hybrid models. The selected methods are composed of modified SMOTE algorithms in 12 (34.3%), standard SMOTE in 10 (28.6%), ADASYN in 2 (5.7%), and AHC-OS in 1 (2.9%). Finally, GAN-based oversampling performed as the best techniques in 3 out of 4 papers (75%)—with GAN [65], WGAN [66], and CWGAN [67].

Pure undersampling techniques perform worse than the other types. Only 3 papers (8.6%) selected undersampling techniques—ENN [45], CDNN [46], KMPP-US [63]. Nevertheless, Tashkandi and Wiese [63] only compared KMPP-US other undersampling techniques, and the relative performance for ENN and CDNN was low—below or equal to 50%. Additionally, Vu et al. [46] achieved similar results with CDNN and SVM-SMOTE (oversampling).

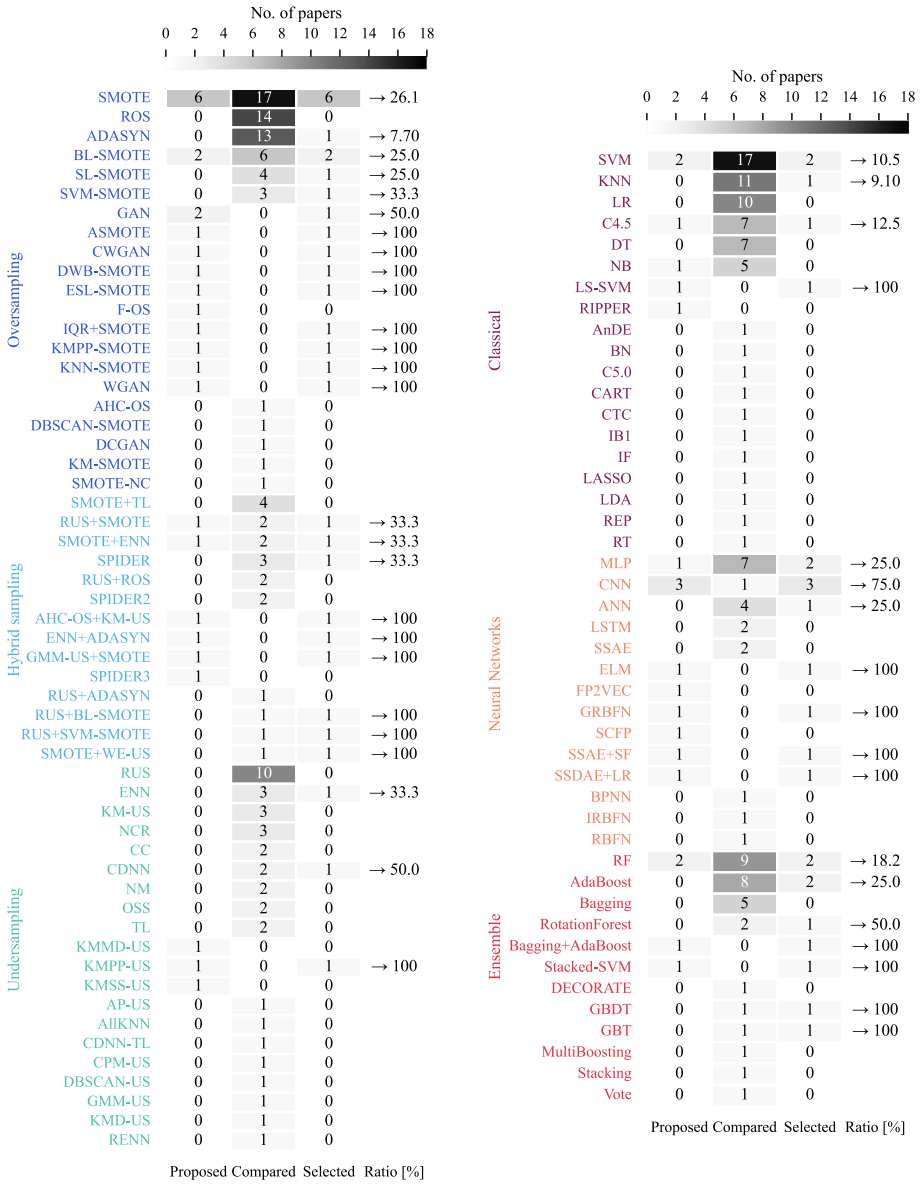

Fig. 4 Quantitative analysis of the reviewed papers proposing, comparing, and selecting: **a** sampling techniques; **b** Machine Learning models

Similarly to Tashkandi and Wiese [63] for undersampling, 15 works [52–62, 65–67, 69] applied or compared only oversampling techniques. This approach accounts for 46.9% of the selected techniques and creates a bias towards oversampling. None of the studies using hybrid sampling used this approach, improving the credibility of their results. Specifically, 6 papers [38–42, 47] selected hybrid sampling techniques after comparing them with the pure oversampling and undersampling techniques composing them.

Three works indicate the need for testing different sampling techniques [46, 49, 70]. More specifically, 3 other works name the need for testing GAN-based oversampling [39, 58, 68], since related studies achieved good results. Reviewed studies also support GAN-based approaches in their conclusions. Liu et al. [65] claim that GAN improves experimental accuracy as the IR increases when compared to other sampling techniques. Gangwar and Ravi [66] assert that WGAN outperforms GAN due to having a better objective function, as well as envision investigating different generator architectures for improving results even more.

Concerning the importance of hybrid sampling, 2 studies affirm that this is the best sampling type for improving the classification of imbalanced data [40, 47]. According to both studies, undersampling alone causes loss of information, while oversampling alone might cause induced bias or overfitting—especially in highly imbalanced datasets.

4.4 FQ3: Are there any studies that use simulation data for preprocessing imbalanced datasets?

There is only one study using simulation data—on power system stability [68]. However, the authors used simulated data for training and testing, not as preprocessing support for real-world data tests.

Even so, some of the domain areas have potentially applicable simulators for synthetic data generation. For instance, electrical and mechanical engineering have fault simulators, and health has exam simulators to this end.

From the 35 studies, 28 (80%) selected solutions based on synthetic oversampling (SMOTE, ADASYN, AHC-OS, and GAN). Thus, using simulation data in suitable domain areas can represent a means for optimizing results and accelerating training time. This acceleration is essential due to the high computational cost for synthetic data generation.

4.5 FQ4: Which Machine Learning models are used in imbalanced data applications?

Similar to preprocessing techniques, the studied works test a wide variety of ML models to improve predictions. From the 35 works, 15 (42.9%) propose a specific model and compare it against alternatives. Conversely, 20 works (57.1%) implement empirical analyses comparing multiple ML models to discuss results and select the best one.

Hence, following the method applied to sampling techniques, this review proposes a taxonomy and a quantitative description of all ML models from the reviewed studies. The taxonomy in Fig. 5 allocates models by their category, dividing into “classical”, “Neural Networks” (NN), and “ensemble”. The figure also indicates the ID of papers applying each model below the corresponding box. Moreover, Fig. 4 details the number of papers for each ML model—grouping by the corresponding category, classifying between “proposed”, “compared”, and “selected”, as well as showing the relative performance.

The distribution of models in Fig. 4 indicates a substantial interest in classical supervised learning models for empirical studies—such as SVM, KNN, LR, DT, and NB. Even so, most

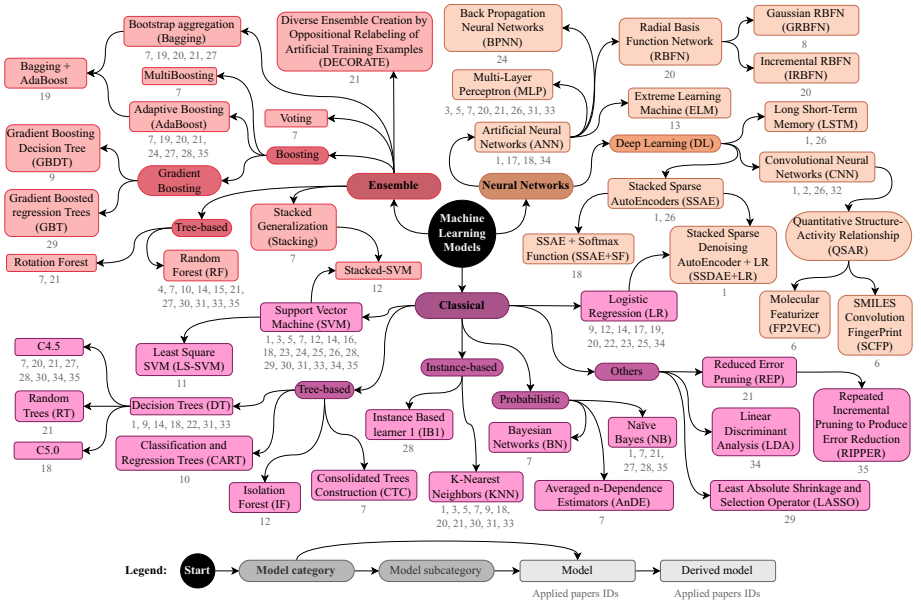

Fig. 5 Taxonomy of Machine Learning models proposed or compared in the reviewed papers by ID

of the proposed ML models involve ANN and CNN optimally configured for the application [57, 58, 61, 65]. This is specially noticeable in works of singularly used models, such as SSDAE+LR [68], SCFP and FP2VEC [39], GRBFN [52], ELM [54], and SSAE+SF [60].

Comparatively, 6 out of 7 studies (85.7%) testing both NN and classical models achieved better performance with NN models [45, 57, 58, 60, 62, 68]. Additionally, 5 NN models achieved the best performance when not compared with classical models [39, 52, 54, 61, 65]. Conversely, 4 classical models achieved the best performance when not tested against NN models [38, 41, 53, 69]—besides the 1 out of 7 studies that did and performed better [48].

Ultimately, ensemble models correspond to 9 (25.7%) of the best performing out of 35 papers. The results in Fig. 4 indicate that RF, AdaBoost, and Bagging are frequently applied—even with preprocessed imbalanced data.

The superiority of NN and ensemble models is noticeable in the studies' conclusions, mentioning the lack of these model categories as a limitation. Incidentally, 4 works expect to apply NN models in future implementations [45, 61, 67, 71]. Additionally, 4 works want to apply ensemble models [47, 49, 59, 71].

Finally, Jiang et al. [61] argue the importance of evaluating the most meaningful metrics for improving imbalanced datasets—since many studies only consider the system's accuracy. Different applications have different priorities. For instance, Haldar et al. [48] focus on improving the sensitivity of the minority class while sufficiently preserving the accuracy in epileptic seizure detection (health). In applications such as disease detection, it is better to guarantee all TPs (diagnoses) possible, even though this creates more FPs.

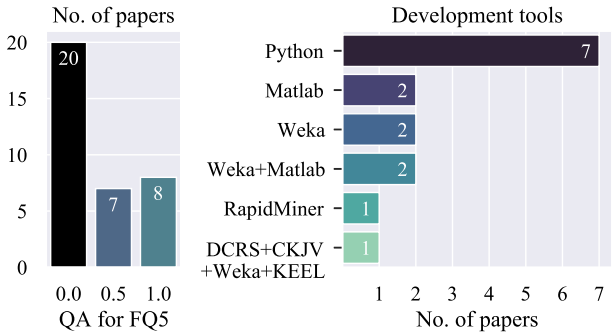

Fig. 6 Development tools of the reviewed applications

4.6 FQ5: Which development tools are used for implementing the proposed solutions?

Figure 6 shows the development tools applied for implementing the studies' solutions. The difficulty in answering this FQ is that most papers do not report any tools used for data processing and ML. These papers account for 20 works (57.1%). The remaining 15 papers report using at least one tool for data preprocessing, ML training, and testing. The completion of this answer—such as programming language, package, and software—corresponds to the QA score for FQ5 in Table 5.

The programming language Python is the most used tool, with ML models through the packages Scikit-learn [66, 71], Keras [58], Tensorflow and Chainer [39]. Additionally, text data applications use natural language processing packages, such as SpaCy [40], NLPNET and NLTK [71]. Other use cases implement sampling techniques through Imbalanced-learn [66] and user-developed scripts [43, 64].

Another programming language applied in the studies is MATLAB. Two studies employ the language for implementing both preprocessing and ML models [60, 62]. In contrast, two studies create a test system with MATLAB in conjunction with standard ML models from the software Weka [49, 72]. Moreover, other solutions use only Weka for all experiments—such as feature selection, sampling, ML training, and testing [52, 65].

Tashkandi and Wiese [63] compared solutions with the software RapidMiner Studio—combining preprocessing, modeling, training, and testing. Finally, Malhotra and Lata [50] created a testing system with the following tools: Data Collection and Reporting System (DCRS) tool for data extraction through GIT repositories; Chidamber and Kemerer Java Metrics (CKJV) tool for object-oriented metrics in Java source codes; Weka for outlier analysis through IQR; Knowledge Extraction based on Evolutionary Learning (KEEL) tool for sampling techniques and ML.

4.7 FQ6: Are there any correlations between domain areas and preprocessing techniques or Machine Learning models?

Generally, the 5 central domain areas and "others"—segmented in Sect. 4.2—applied distinctive sampling techniques and ML models in their solutions. Figure 7 details the number of sampling techniques and ML models selected by the authors of at least one paper within the corresponding domain areas. Additionally, studies which did not select and clearly indicate

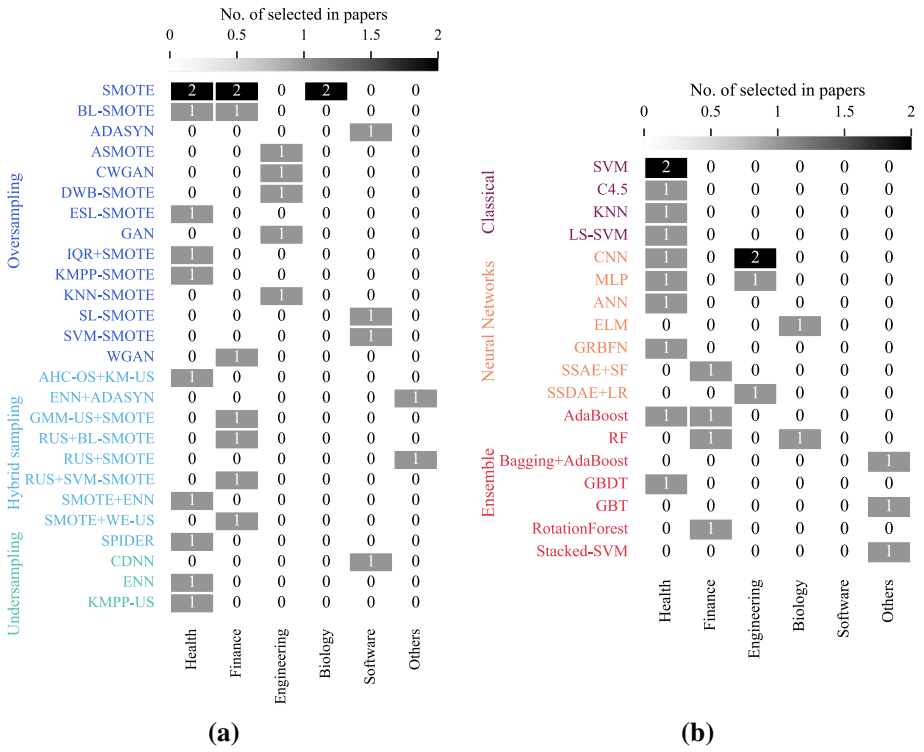

Fig. 7 Quantitative analysis in different domain areas for the selection of: **a** sampling techniques; **b** Machine Learning models

at least one best performing method for the application have not been accounted for—such as ML models in software (Fig. 7b).

Studies on health applied the most diverse methods, potentially due to the substantial proportion of works (34.3%). This domain is the only one applying classical ML models. Additionally, health is the only domain selecting pure undersampling techniques—apart from one study on software [46].

Finance, the second most prevalent domain (22.9%), splits between using oversampling and hybrid sampling techniques. However, for ML categories, 75% of the works indicate a preference for ensemble models. In contrast, one work implements a specialized DL model for bankruptcy prediction (SSDAE +SF)—although it does not compare results with ensemble models [60].

Engineering studies (14.3%) selected an unanimous combination of methods: oversampling and NN models. This domain has all applications related to fault detection—generally suffering from high IR and benefiting from oversampling techniques.

Similarly to engineering, biology studies (8.6%) also selected only oversampling—through SMOTE. Additionally, two studies split ML between NN, with ELM, and ensemble, with RF.

Software studies did not select any best performing ML model. However, the three studies (8.6%) achieved their best results through oversampling. One implementation, by Vu et al. [46], points similar performance between SMOTE-SVM and CDNN (undersampling)—in

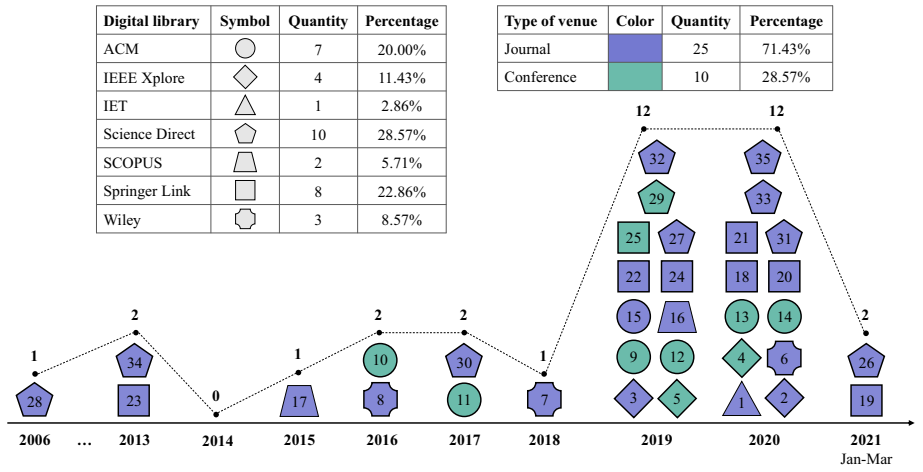

Fig. 8 Reviewed studies per year by digital library and type of venue

less processing time with the latter. Finally, all studies in other areas (11.4%) indicate better performance through hybrid sampling and ensemble models.

As pointed in Sects. 4.3 and 4.5 (FQ2 and FQ4), undersampling techniques and classical models obtain generally worse results than other methods. These studies have only been selected in the bigger sample of studies from the health domain. Therefore, better results should be expected in all domain areas by implementing solutions with combinations of oversampling or hybrid sampling with NN or ensemble models.

4.8 SQ1: How has the quantity of studies evolved?

Figure 8 shows the yearly publication of selected papers by the originated digital library and type of venue. The research was performed at the beginning of April 2021 without any date restraint.

The results indicate a growing interest in data level preprocessing techniques for ML in imbalanced data applications, especially since 2019. Worth noting that EC7 and EC9 filtered out some algorithmic level techniques and solutions for multiple applications—creating a gap of representative works between 2007 and 2012. However, the authors found that papers using these solutions followed a similar pattern of growing interest, presented in Fig. 8.

4.9 SQ2: Where have the studies been published?

The representative works selection integrates 35 publications. This selection shows that 25 journal publications correspond to 71.4%, and 10 conference publications account for 28.6% of the studies reviewed in this paper. Figure 8 indicates the type of venue of these studies by color.

The search process collected most selected papers through the digital libraries ACM, Science Direct, and Springer Link, where each accounts for at least 20% of the results. Additionally, only the journal "Artificial Intelligence in Medicine" has 2 works—one from 2006 [38], and the other from 2020 [69].

Table 7 Lessons learned by answering the research questions

RQ#	Lessons learned
FQ1	There are 5 central domain areas in imbalanced data applications: health (34.3%), finance (22.9%), engineering (14.3%), biology (8.6%), and software (8.6%). These areas have good references for new applications. New domains have the potential to be explored
FQ2	The studies applied 55 different sampling techniques—oversampling (55.5%), undersampling (27.4%), and hybrid sampling (17.1%). Oversampling techniques achieved the best performance among the existing types, whereas hybrid sampling techniques performed better relatively (ratio of selected within tested studies)
FQ3	None of the studies used simulation as a means for optimizing synthetic data generation and accelerating training time in oversampling. This technology could optimize results and reduce computational costs in domains such as engineering and health
FQ4	The studies applied 45 different ML models—classical (54%), ensemble (24.8%), and NN (21.2%). NN models achieved the best performance overall and relative to tested studies, with ensemble models as a close second
FQ5	There are 3 recurrent development tools within the studies: Python, MATLAB, and Weka. These tools have both sampling techniques and ML models already implemented as resources
FQ6	Domain areas selected distinctive sampling techniques and ML models—especially in health. However, there is a clear preference for oversampling in engineering, biology, and software, while finance splits between oversampling and hybrid sampling. For ML, engineering selected only NN models, and finance selected mostly ensemble models. Other domains did not have a clear categorical preference
SQ1	There is a growing research interest in the subject, especially since 2019
SQ2	The 35 reviewed studies show a prevalence of journal publications, with 25 works (71.4%), while the remaining 10 are from conferences. The digital libraries ACM, Science Direct, and Springer Link account for at least 20% of the results individually

5 Conclusion

This paper applied a systematic mapping study to review current and effective data level preprocessing techniques and ML models in imbalanced data applications. After an eight-step filtering process, the selection of the representative works culminated in 35 papers. The results section presents two taxonomies and quantitative classifications of proposed, compared, and selected preprocessing techniques and ML models.

Overall, research studies mainly focus on applying standard or modified clustering-based sampling techniques for balancing data. Specifically, oversampling is the most common and also the best performing type of sampling, proportionally. Relatively, however, hybrid sampling techniques can potentially surpass oversampling if future studies implement them.

Classical ML models such as SVM, KNN, and LR still are the most frequent. Nevertheless, recent studies show an increase in NN models—from simple ANNs, like MLP, to complex DL models. The results indicate that well configured NN models tend to achieve better results than classical models. Additionally, ensemble learning models also show promising results.

Ultimately, the results found in this systematic mapping study indicate that future works may explore the usage of simulation-based oversampling for balancing data in ML applications. Moreover, a solution with hybrid sampling mixed with NN or ensemble learning models can potentially achieve favorable results. Table 7 compiles the highlights from RQs' answers.

The lack of analyzable dataset characteristics is a limiting factor for this study. In future literature reviews, the authors suggest the addition of an EC if studies do not present the information of interest. An alternate study could be performed by reviewing papers with well-known prefixed datasets from different domain areas.

Acknowledgements This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001, and by the National Council for Scientific and Technological Development (CNPq). We would also like to thank the University of Vale do Rio dos Sinos (Unisinos).

Author Contributions All authors contributed to the study, read and approved the final manuscript. The list below describes the CRediT (Contributor Roles Taxonomy) by author: VWV: Conceptualization, Methodology, Formal analysis, Writing—Original Draft; JASA: Writing—Review and Editing; RSC: Writing—Review and Editing; PRSP: Writing—Review and Editing, Supervision; JLVB: Writing—Review and Editing, Supervision.

Funding This study was financed in part by the following Brazilian federal organizations. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)—Finance Code 001; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)—Award Number 306395/2017-7.

Availability of data and materials Not applicable.

Declarations

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Code availability Not applicable.

References

1. Zhang C, Zhou Y, Deng Y (2019) VCOS: a novel synergistic oversampling algorithm in binary imbalance classification. IEEE Access 7:145435–145443. https://doi.org/10.1109/ACCESS.2019.2945034
2. Fotouhi S, Asadi S, Kattan MW (2019) A comprehensive data level analysis for cancer diagnosis on imbalanced data. J Biomed Inform 90:103089. https://doi.org/10.1016/j.jbi.2018.12.003
3. Rekha G, Krishna Reddy V, Tyagi AK (2020) An Earth mover's distance-based undersampling approach for handling class-imbalanced data. Int J Intell Inf Database Syst 13(2–4):376–392. https://doi.org/10.1504/IJIIDS.2020.109463
4. Wong GY, Leung FHF, Ling SH (2014) A novel evolutionary preprocessing method based on oversampling and under-sampling for imbalanced datasets. In: IECON 2013—39th annual conference of the IEEE industrial electronics society, pp. 2354–2359. IEEE, Vienna, Austria. https://doi.org/10.1109/IECON.2013.6699499
5. Zhang J, Cui X, Li J, Wang R (2017) Imbalanced classification of mental workload using a cost-sensitive majority weighted minority oversampling strategy. Cogn Technol Work 19(4):633–653. https://doi.org/10.1007/s10111-017-0447-x
6. Dong Y, Wang X (2011) A new over-sampling approach: random-SMOTE for learning from imbalanced data sets. In: KSEM 2011: 5th international conference on knowledge science, engineering and management, pp. 343–352. Springer, Irvine, USA. https://doi.org/10.1007/978-3-642-25975-3_30

7. Zhao SX, Wang XL, Yue QS (2020) A novel mixed sampling algorithm for imbalanced data based on XGBoost. In: CWSN 2020: 14th China conference on wireless sensor networks, pp 181–196. Springer, Dunhuang, China. https://doi.org/10.1007/978-981-33-4214-9_14
8. Kaur H, Pannu HS, Malhi AK (2019) A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Comput Surv 52(4):1–36. https://doi.org/10.1145/3343440
9. Felix EA, Lee SP (2019) Systematic literature review of preprocessing techniques for imbalanced data. IET Softw 13(6):479–496. https://doi.org/10.1049/iet-sen.2018.5193
10. Spelmen VS, Porkodi R (2018) A review on handling imbalanced data. In: 2018 international conference on current trends towards converging technologies (ICCTCT), pp 1–11. IEEE, Coimbatore, India. https://doi.org/10.1109/ICCTCT.2018.8551020
11. Susan S, Kumar A (2020) The balancing trick: optimized sampling of imbalanced datasets—a brief survey of the recent State of the Art. Eng Rep 3(4):1–24. https://doi.org/10.1002/eng2.12298
12. Shakeel F, Sabhitha AS, Sharma S (2017) Exploratory review on class imbalance problem: an overview. In: 2017 8th international conference on computing, communication and networking technologies (ICCCNT), pp 1–8. IEEE, Delhi, India. https://doi.org/10.1109/ICCCNT.2017.8204150
13. Johnson JM, Khoshgoftaar TM (2019) Survey on deep learning with class imbalance. J Big Data 6:1–54. https://doi.org/10.1186/s40537-019-0192-5
14. Li Q, Mao Y (2014) A review of boosting methods for imbalanced data classification. Pattern Anal Appl 17:679–693. https://doi.org/10.1007/s10044-014-0392-8
15. Buda M, Maki A, Mazurowski MA (2018) A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw 106:249–259 arXiv:1710.05381. https://doi.org/10.1016/j.neunet.2018.07.011
16. Bhatore S, Mohan L, Reddy YR (2020) Machine learning techniques for credit risk evaluation: a systematic literature review. J Bank Financ Technol 4(1):111–138. https://doi.org/10.1007/s42786-020-00020-3
17. Sirsat MS, Fermé E, Câmara J (2020) Machine learning for brain stroke: a review. J Stroke Cerebrovasc Dis 29(10):105162. https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105162
18. Thanoun MY, Yaseen MT (2020) A comparative study of Parkinson disease diagnosis in machine learning. In: ICAAI 2020: 2020 the 4th international conference on advances in artificial intelligence, pp 23–28. ACM, New York, USA. https://doi.org/10.1145/3441417.3441425
19. Chugh G, Kumar S, Singh N (2021) Survey on machine learning and deep learning applications in breast cancer diagnosis. Cogn Comput. https://doi.org/10.1007/s12559-020-09813-6
20. Ishtiaq U, Abdul Kareem S, Abdullah ERMF, Mujtaba G, Jahangir R, Ghafoor HY (2020) Diabetic retinopathy detection through artificial intelligent techniques: a review and open issues. Multimed Tools Appl 79:15209–15252. https://doi.org/10.1007/s11042-018-7044-8
21. Hu Z, Tang J, Wang Z, Zhang K, Zhang L, Sun Q (2018) Deep learning for image-based cancer detection and diagnosis—a survey. Pattern Recogn 83:134–149. https://doi.org/10.1016/j.patcog.2018.05.014
22. Benhar H, Idri A, Fernández-Alemán JL (2020) Data preprocessing for heart disease classification: a systematic literature review. Comput Methods Programs Biomed 195:105635. https://doi.org/10.1016/j.cmpb.2020.105635
23. Idri A, Benhar H, Fernández-Alemán JL, Kadi I (2018) A systematic map of medical data preprocessing in knowledge discovery. Comput Methods Programs Biomed 162:69–85. https://doi.org/10.1016/j.cmpb.2018.05.007
24. Lei Y, Yang B, Jiang X, Jia F, Li N, Nandi AK (2020) Applications of machine learning to machine fault diagnosis: a review and roadmap. Mech Syst Signal Process 138:106587. https://doi.org/10.1016/j.ymssp.2019.106587
25. Zhang T, Chen J, Li F, Zhang K, Lv H, He S, Xu E (2021) Intelligent fault diagnosis of machines with small and imbalanced data: a state-of-the-art review and possible extensions. ISA Trans. https://doi.org/10.1016/j.isatra.2021.02.042
26. Amarasinghe T, Aponso A, Krishnarajah N (2018) Critical analysis of machine learning based approaches for fraud detection in financial transactions. In: ICMLT'18: Proceedings of the 2018 international conference on machine learning technologies, pp 12–17. ACM, New York, USA. https://doi.org/10.1145/3231884.3231894
27. Priscilla CV, Prabha DP (2019) Credit card fraud detection: a systematic review. In: Proceedings of the first international conference on innovative computing and cutting-edge technologies (ICICCT 2019), pp 290–303. Springer, Istanbul, Turkey. https://doi.org/10.1007/978-3-030-38501-9_29
28. Li Z, Jing XY, Zhu X (2018) Progress on approaches to software defect prediction. IET Softw 12(3):161–175. https://doi.org/10.1049/iet-sen.2017.0148
29. Pandey SK, Mishra RB, Tripathi AK (2021) Machine learning based methods for software fault prediction: a survey. Expert Syst Appl 172:114595. https://doi.org/10.1016/j.eswa.2021.114595

30. Malhotra R (2015) A systematic review of machine learning techniques for software fault prediction. Appl Soft Comput 27:504–518. https://doi.org/10.1016/j.asoc.2014.11.023
31. Guzella TS, Caminhas WM (2009) A review of machine learning approaches to Spam filtering. Expert Syst Appl 36(7):10206–10222. https://doi.org/10.1016/j.eswa.2009.02.037
32. Kitchenham B, Pretorius R, Budgen D, Brereton OP, Turner M, Niazi M, Linkman S (2010) Systematic literature reviews in software engineering—a tertiary study. Inf Softw Technol 52(8):792–805. https://doi.org/10.1016/j.infsof.2010.03.006
33. Cooper ID (2016) What is a "mapping study?". J Med Libr Assoc 104(1):76–78. https://doi.org/10.3163/1536-5050.104.1.013
34. Petersen K, Vakkalanka S, Kuzniarz L (2015) Guidelines for conducting systematic mapping studies in software engineering: an update. Inf Softw Technol 64:1–18. https://doi.org/10.1016/j.infsof.2015.03.007
35. De Almeida LG, Souza ADD, Kuehne BT, Gomes OSM (2020) Data analysis techniques in vehicle communication networks: systematic mapping of literature. IEEE Access 8:199503–199512. https://doi.org/10.1109/access.2020.3034588
36. Silva RDA, Braga RTV (2020) Simulating systems-of-systems with agent-based modeling: a systematic literature review. IEEE Syst J 14(3):3609–3617. https://doi.org/10.1109/JSYST.2020.2980896
37. Keshav S (2007) How to read a paper. ACM SIGCOMM Comput Commun Rev 37(3):83–84. https://doi.org/10.1145/1273445.1273458
38. Cohen G, Hilario M, Sax H, Hugonnet S, Geissbuhler A (2006) Learning from imbalanced data in surveillance of nosocomial infection. Artif Intell Med 37(1):7–18. https://doi.org/10.1016/j.artmed.2005.03.002
39. Lee YO, Kim YJ (2020) The effect of resampling on data-imbalanced conditions for prediction towards nuclear receptor profiling using deep learning. Mol Inf 39(8):1900131. https://doi.org/10.1002/minf.201900131
40. Mahadevan A, Arock M (2021) A class imbalance-aware review rating prediction using hybrid sampling and ensemble learning. Multimed Tools Appl 80(5):6911–6938. https://doi.org/10.1007/s11042-020-10024-2
41. Rustam Z, Utami DA, Hidayat R, Pandelaki J, Nugroho WA (2019) Hybrid preprocessing method for support vector machine for classification of imbalanced cerebral infarction datasets. Int J Adv Sci Eng Inf Technol 9(2):685–691. https://doi.org/10.18517/ijaseit.9.2.8615
42. Chang Q, Lin S, Liu X (2019) Stacked-SVM: a dynamic SVM framework for telephone fraud identification from imbalanced CDRs. In: ACAI 2019: proceedings of the 2019 2nd international conference on algorithms, computing and artificial intelligence, vol 9, pp 112–120. ACM, New York, USA. https://doi.org/10.1145/3377713.3377735
43. Han X, Cui R, Lan Y, Kang Y, Deng J, Jia N (2019) A Gaussian mixture model based combined resampling algorithm for classification of imbalanced credit data sets. Int J Mach Learn Cybern 10(12):3687–3699. https://doi.org/10.1007/s13042-019-00953-2
44. Marqués AI, García V, Sánchez JS (2013) On the suitability of resampling techniques for the class imbalance problem in credit scoring. J Oper Res Soc 64(7):1060–1070. https://doi.org/10.1057/jors.2012.120
45. Pereira RM, Bertolini D, Teixeira LO, Silla CN, Costa YMG (2020) COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. Comput Methods Programs Biomed 194:105532 arXiv:2004.05835. https://doi.org/10.1016/j.cmpb.2020.105532
46. Vu L, Van Tra D, Nguyen QU (2016) Learning from imbalanced data for encrypted traffic identification problem. In: SoICT'16: proceedings of the seventh symposium on information and communication technology, pp 147–152. ACM, New York, USA. https://doi.org/10.1145/3011077.3011132
47. Shamsudin H, Yusof UK, Jayalakshmi A, Akmal Khalid MN (2020) Combining oversampling and undersampling techniques for imbalanced classification: a comparative study using credit card fraudulent transaction dataset. In: 2020 IEEE 16th international conference on control and automation (ICCA), pp 803–808. IEEE, Singapore. https://doi.org/10.1109/ICCA51439.2020.9264517
48. Haldar S, Mukherjee R, Chakraborty P, Banerjee S, Chaudhury S, Chatterjee S (2019) Improved epilepsy detection method by addressing class imbalance problem. In: 2018 IEEE 9th annual information technology, electronics and mobile communication conference (IEMCON), pp 934–939. IEEE, Vancouver, BC, Canada. https://doi.org/10.1109/IEMCON.2018.8614826
49. Malhotra R, Kamal S (2019) An empirical study to investigate oversampling methods for improving software defect prediction using imbalanced data. Neurocomputing 343:120–140. https://doi.org/10.1016/j.neucom.2018.04.090
50. Malhotra R, Lata K (2020) An empirical study on predictability of software maintainability using imbalanced data. Softw Qual J 28(4):1581–1614. https://doi.org/10.1007/s11219-020-09525-y

51. Ma J, Afolabi DO, Ren J, Zhen A (2019) Predicting seminal quality via imbalanced learning with evolutionary safe-level synthetic minority over-sampling technique. Cogn Comput. https://doi.org/10.1007/s12559-019-09657-9
52. Yan S, Qian W, Guan Y, Zheng B (2016) Improving lung cancer prognosis assessment by incorporating synthetic minority oversampling technique and score fusion method. Med Phys 43(6):2694–2703. https://doi.org/10.1118/1.4948499
53. Purnami SW, Trapsilasiwi RK (2017) SMOTE-least square support vector machine for classification of multiclass imbalanced data. In: ICMLC 2017: proceedings of the 9th international conference on machine learning and computing, pp 107–111. ACM, New York, USA. https://doi.org/10.1145/3055635.3056581
54. Dewi C, Firdaus Mahmudy W, Arifando R, Kusuma Arbawa Y, Labique Ahmadie B, Labique B (2020) Improve performance of extreme learning machine in classification of patchouli varieties with imbalanced class. In: SIET'20: proceedings of the 5th international conference on sustainable information engineering and technology, pp 16–22. ACM, New York, USA. https://doi.org/10.1145/3427423.3427424
55. Zhang X, Lin X, Zhao J, Huang Q, Xu X (2019) Efficiently predicting hot spots in PPIs by combining random forest and synthetic minority over-sampling technique. IEEE/ACM Trans Comput Biol Bioinf 16(3):774–781. https://doi.org/10.1109/TCBB.2018.2871674
56. Gicić A, Subasi A (2018) Credit scoring for a microcredit data set using the synthetic minority oversampling technique and ensemble classifiers. Expert Syst 36(2):1–22. https://doi.org/10.1111/exsy.12363
57. Tra V, Duong BP, Kim JM (2019) Improving diagnostic performance of a power transformer using an adaptive over-sampling method for imbalanced data. IEEE Trans Dielectr Electr Insul 26(4):1325–1333. https://doi.org/10.1109/TDEI.2019.008034
58. Jiang N, Li N (2021) A wind turbine frequent principal fault detection and localization approach with imbalanced data using an improved synthetic oversampling technique. Int J Electr Power Energy Syst 126 Part A:106595. https://doi.org/10.1016/j.ijepes.2020.106595
59. Faris H, Abukhurma R, Almanaseer W, Saadeh M, Mora AM, Castillo PA, Aljarah I (2020) Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: a case from the Spanish market. Prog Artif Intell 9(1):31–53. https://doi.org/10.1007/s13748-019-00197-9
60. Smiti S, Soui M (2020) Bankruptcy prediction using deep learning approach based on borderline SMOTE. Inf Syst Front 22(5):1067–1083. https://doi.org/10.1007/s10796-020-10031-6
61. Jiang J, Zhang H, Pi D, Dai C (2019) A novel multi-module neural network system for imbalanced heartbeats classification. Expert Syst Appl X 1:100003. https://doi.org/10.1016/j.eswax.2019.100003
62. Santos MS, Abreu PH, García-Laencina PJ, Simão A, Carvalho A (2015) A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. J Biomed Inform 58:49–59. https://doi.org/10.1016/j.jbi.2015.09.012
63. Tashkandi A, Wiese L (2019) A hybrid machine learning approach for improving mortality risk prediction on imbalanced data. In: iiWAS2019: proceedings of the 21st international conference on information integration and web-based applications and services, pp 83–92. ACM, New York, USA. https://doi.org/10.1145/3366030.3366040
64. Zhou Q, Sun B, Song Y, Li S (2020) K-means clustering based undersampling for lower back pain data. In: ICBDT 2020: proceedings of the 2020 3rd international conference on big data technologies, pp 53–57. ACM, New York, USA. https://doi.org/10.1145/3422713.3422725
65. Liu Q, Ma G, Cheng C (2020) Data fusion generative adversarial network for multi-class imbalanced fault diagnosis of rotating machinery. IEEE Access 8:70111–70124. https://doi.org/10.1109/ACCESS.2020.2986356
66. Gangwar AK, Ravi V (2019) WiP: generative adversarial network for oversampling data in credit card fraud detection. In: ICISS 2019: 15th international conference on information systems security, vol 11952, pp 123–134. Springer, Hyderabad, India. https://doi.org/10.1007/978-3-030-36945-3
67. Yan K, Huang J, Shen W, Ji Z (2020) Unsupervised learning for fault detection and diagnosis of air handling units. Energy Build 210:109689. https://doi.org/10.1016/j.enbuild.2019.109689
68. Wang H, Ye W (2020) Transient stability evaluation model based on SSDAE with imbalanced correction. IET Gener Transm Distrib 14(11):2209–2216. https://doi.org/10.1049/iet-gtd.2019.1388
69. Nnamoko N, Korkontzelos I (2020) Efficient treatment of outliers and class imbalance for diabetes prediction. Artif Intell Med 104:101815. https://doi.org/10.1016/j.artmed.2020.101815
70. Liu S, Wang Y, Zhang J, Chen C, Xiang Y (2017) Addressing the class imbalance problem in Twitter spam detection using ensemble learning. Comput Secur 69:35–49. https://doi.org/10.1016/j.cose.2016.12.004
71. Filho AH, Concatto F, Nau J, Prado HAD, Imhof DO, Ferneda E (2019) Imbalanced learning techniques for improving the performance of statistical models in automated essay scoring. In: Knowledge-based and intelligent information & engineering systems: proceedings of the 23rd international conference

KES2019, vol 159, pp 764–773. Elsevier B.V., Budapest, Hungary. https://doi.org/10.1016/j.procs.2019.09.235
72. Zhou L (2013) Performance of corporate bankruptcy prediction models on imbalanced dataset: the effect of sampling methods. Knowl Based Syst 41:16–25. https://doi.org/10.1016/j.knosys.2012.12.007

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Vitor Werner de Vargas is a Master's student and researcher in Applied Computing at the University of Vale do Rio dos Sinos (Unisinos). He received his Bachelor's degree in Electrical Engineering from Unisinos in 2020. His research interests include Data Analysis, Machine Learning, Automated Systems, Smart Grids, and Power Systems.

Jorge Arthur Schneider Aranda is Researcher at the University of Vale do Rio dos Sinos (UNISINOS). Doctoral candidate in Applied Computing and Master's degree in applied computing at the University of Vale dos Sinos, Graduated in Computer Science at the Feevale University. He has experience in Machine Learning, Networks, Internet of Things, E-Health and Ubiquitous Computing.

Ricardo dos Santos Costa is a master's student and researcher in electrical engineering at the University of Vale do Rio dos Sinos (Unisinos). Graduated in Mathematics in 2011 and Bachelor in Electrical Engineering from Unisinos in 2018. His research area includes Data Analysis, Artificial Intelligence, Smart Grids and Power Systems.

Paulo Ricardo da Silva Pereira is a professor at University of Vale do Rio dos Sinos (Unisinos). He works with Power Systems since 1998, at companies Certaja, RGE, and CEEE. He received his M.Sc. and Ph.D. degrees in Electrical Engineering from Federal University of Santa Maria, Brazil, in 2009 and 2014, respectively. His research interests include Power Systems, Renewable and Distributed Energy Resources, Power Quality, Embedded Systems, and Automation.

Jorge Luis Victória Barbosa received M.Sc. and Ph.D. in computer science from the Federal University of Rio Grande do Sul, Brazil. He conducted post-doctoral studies at Sungkyunkwan University (South Korea) and University of California Irvine (USA). Jorge is a full professor of the University of Vale do Rio dos Sinos (UNISINOS), Brazil.