A Multi-Dimensional Survey on Learning from Imbalanced Data

Leonidas Akritidis and Panayiotis Bozanis

Abstract The problem of data imbalance in machine learning is related to the uneven distribution of the training examples to the involved classes. Nowadays, a large number of research fields and applications suffer from class imbalance, including Cybersecurity, Bioinformatics, Natural Language Processing, management of multimedia content, and so on. Directly training machine/deep learning classifiers with such data has been proved quite problematic, because the generated models become strongly biased towards the majority class. Unable to learn the minority classes effectively, the accuracy of these "imbalanced" models degrades rapidly. Given the importance of the problem, numerous researchers have introduced innovative, state-of-the-art approaches with the aim of addressing it. In this chapter, we present a survey on the most significant advances in the area, by adopting a multi-dimensional categorization approach. Specifically, we classify the relevant works according to: i) the application field they focus on, ii) the methods they introduce to mitigate class imbalance, and iii) the classification models they utilize to evaluate the introduced algorithms. Additionally, we cover the state-of-the-art overviews in a systematic manner and we describe the proposed methods and their achieved results.

1 Introduction

The problem of data imbalance occurs when the input samples of a dataset are unequally distributed across its classes. For example, a collection of 100 images, of which 80 illustrate dogs and the rest of them depict other animals, constitutes an

Leonidas Akritidis
Department of Science and Technology, International Hellenic University, 14th km Thessaloniki – N. Moudania, 57001, Thermi, Thessaloniki, Greece, e-mail: lakritidis@ihu.gr

Panayiotis Bozanis
Department of Science and Technology, International Hellenic University, 14th km Thessaloniki – N. Moudania, 57001, Thermi, Thessaloniki, Greece e-mail: pbozanis@ihu.gr

imbalanced dataset. The class that contains the majority of the samples is called the majority class (the dog class in our example), whereas the other classes are known as minority classes. In extreme situations, the majority class vastly outnumbers the minority classes (a ratio of say 1000:1 or more), so the problem is often referred as extreme data imbalance.

Imbalanced datasets are common in a wide range of applications, including computer vision, image processing, intrusion, malware and fraud detection, NLP, sentiment analysis, bioinformatics, etc. [53]. In these fields, most real-world data collections are imbalanced to one extent or another. Unfortunately, such collections introduce significant problems in classification tasks, because the underlying models have an enormous difficulty in learning the minority class. In particular, the classifiers become biased towards the majority class and their (balanced) accuracy degrades significantly.

Due to the severe consequences of data imbalance in classification performance, a large number of research studies have been conducted with the aim of confronting it. During our research, we discovered that the related articles have a three-dimensional nature. The first dimension concerns the application field that the article focuses on. For example, there are numerous works attempting to confront the problem from the perspective of Cybersecurity (e.g., intrusion detection or credit card fraud detection), whereas others emphasize on Bioinformatics (e.g., protein or microRNA classification) or Natural Language Processing (e.g., sentiment analysis).

The second dimension is related to the method that is employed to mitigate the imbalanced class distribution. In this context, multiple algorithms have been proposed either for enriching the minority classes with valuable samples (oversampling), or for removing samples from the majority class (undersampling).

A third category includes the hybrid algorithms that apply both oversampling and undersampling together. Noticeably, the relevant works can be further classified according to the nature of their proposed method. Hence, in the literature we encounter effective oversampling solutions that employ deep generative models (like GANs), nearest neighbor techniques (e.g. SMOTE), and feature engineering approaches.

Although the problem of data imbalance is of particular importance for all data mining and machine learning applications, the articles encountered in the literature mainly focus on classification tasks. A common methodology adopted by many relevant works is to examine the effectiveness of a data imbalance solution in combination with a variety of classifiers. Hence, usually the authors first describe a data preprocessing technique, and then they evaluate it by applying a set of classifiers on several test datasets. To cover this aspect, we present a third way of categorizing the relevant literature. We call this organization as classifier-based, because it is based on the classification models that have been devised or employed to evaluate a data imbalance algorithm.

The rest of the chapter is organized as follows: Section 2 describes our research methodology and introduces its triadic nature. In the sequel, Sections 3, 4 and 5 present the state-of-the-art works from an application-based, technique-based, and classifier-based perspective, respectively. Finally, Section 7 concludes the overview with important findings and key observations.

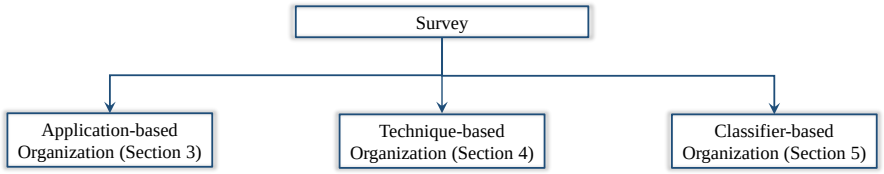

Fig. 1 The triadic (or 3-dimensional) nature of our survey: We examine the relevant works from the perspective of the applications they focus on, the techniques they introduce and the classifiers they employ.

2 Methodology and Organization

The articles presented in this chapter have been retrieved by using one academic search engine (Google Scholar[1]), and four scientific digital libraries: SpringerLink[2], ACM library[3], IEEE Xplore[4], and ScienceDirect[5]. The queries that we submitted to these services were formed according to the application field, or the approach that was adopted to confront data imbalance. We provide more details on how the queries were crafted in the list below.

The presentation is based on a three-dimensional approach. Each dimension represents a different perspective from which each paper can be viewed, or a different category where each study belongs (Figure 1). Specifically:

- Application-based organization: The studies are categorized on the basis of the application(s) they focus on. They were collected by submitting application specific queries to the aforementioned services, concatenated with the phrase *"imbalanced data"*. For example, to retrieve articles that studied the problem from the perspective of malware detection, we submitted the query *"malware detection imbalanced data"*. The hierarchical structure of Figure 2 implies the formation of 28 such queries. The most important results that were retrieved by this procedure are discussed in Section 3.
- Technique-based organization: The papers are presented according to the techniques that they introduce (or apply) to confront data imbalance. Notice that a significant number of relevant works in the area examine the general-case version of the problem; that is, without considering any particular application. Therefore, the organization of Section 4 includes both generic and application-specific solutions. Similarly to the previous procedure, we formed multiple targeted queries to locate relevant articles. For example, we submitted the query *"GAN data imbalance"* to retrieve works that utilized or adapted Generative Adversarial Networks to oversample the minority class.

[1] https://scholar.google.com

[2] https://link.springer.com/

[3] https://dl.acm.org/

[4] https://ieeexplore.ieee.org/Xplore/home.jsp

[5] https://www.sciencedirect.com/

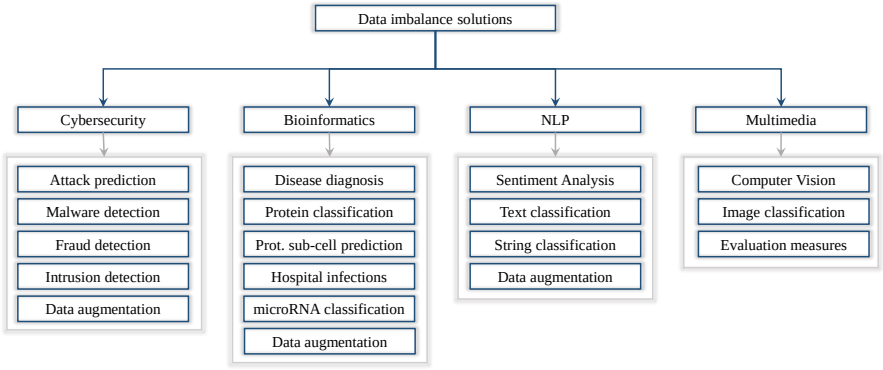

Fig. 2 Application-wise categorization of methods for handling data imbalance.

- Classifier-based organization: Data imbalance has been extensively studied in the context of classification, either from a generic or an application-specific point of view. Several researchers have adapted well-established models to address the problem, whereas others utilized a variety of classifiers to evaluate their techniques on test datasets. A brief overview of these classifiers is presented in the third part of this chapter.

3 Application-based Organization

As mentioned earlier, the problem of data imbalance concerns a large number of applications from diverse scientific, enterprise and industrial fields. In the following subsections we present the most indicative applications and we refer to studies that examined the problem of data imbalance from the perspective of these applications. The works cited below are summarized per category in Table 1. A visual representation of the aforementioned categorization is provided in Figure 2.

3.1 Cybersecurity

Cybersecurity is a broad field that studies security issues for software and computer systems. It includes various sub-areas that deal with malware and virus detection, fraud and spam detection, intrusion detection, and so forth. The most recent solutions that confront these problems rely on machine learning models that predict whether an action is normal or malicious. Nonetheless, training such models is a challenging task because the utilized datasets usually include highly imbalanced samples. For instance, in intrusion detection scenarios, the training examples that represent legitimate actions vastly outnumber those that represent attack actions. Similarly, in

Table 1 Data imbalance solutions categorized by application field.

Application field	Application	Papers
Cybersecurity	Attack prediction	[6], [102]
	Malware detection	[25], [76]
	Fraud detection	[36], [72], [86]
	Intrusion detection	[3], [9], [11], [28], [37], [63], [118]
	Data augmentation	[109]
Bioinformatics and Medical Sciences	Disease diagnosis	[10], [58], [83], [100], [116]
	Protein Classification	[93], [123]
	Protein sub-cellular prediction	[101]
	Hospital infections	[26]
	microRNA classification	[16]
	Data augmentation	[29], [121]
Natural Language Processing	Sentiment Analysis	[44], [60], [77], [104]
	Text classification	[75], [79], [80], [112], [94]
	String classification	[19]
	Data augmentation	[5], [70]
Multimedia	Computer vision	[42], [87], [113]
	Image classification	[51], [54], [62], [74], [82], [85]
	Evaluation measures	[106]

anti-malware applications the infected files are much fewer than the harmless files. In both cases, training models without confronting the lurking class imbalance leads to biased classifiers with degraded performance.

To address the problem in question, Wang et al. introduced an oversampling model with an attention-based mechanism, called AOPL [102]. The attention mechanism is employed to decrease the redundancy during the generation of the artificial attack examples. The experimental evaluation of AOPL demonstrated its robustness against several deep learning models on four real datasets.

More recently, Akash et al. [6] presented a work that employed the Synthetic Minority Oversampling Technique (SMOTE), Borderline SMOTE and Adaptive Synthetic Sampling (ADASYN) to tackle class imbalance during attack prediction in IoT devices. Interestingly, the authors concluded that none of these methods had a significant impact on the predictive ability of cyber attack detection. On a similar fashion, Wheelus et al. evaluated four preprocessing methods for mitigating class imbalance in the UNSW-NB15 Cybersecurity dataset [109]. They found that only the Bagging technique consistently improved performance; SMOTE had a positive effect only in combination with 3 out of the 5 attested classifiers.

Regarding malware detection, Chen et el. combined several machine learning models with network traffic analysis techniques to recognize malicious applications in mobile devices [25]. To overcome the problem of imbalance in the traffic data, the authors combined SMOTE with Support Vector Machines (SVM), cost-sensitive SVMs, and C4.5 cost-sensitive methods. Another relatively recent malware detection method for mobile devices was proposed by Oak et al. [76]. In that paper, the authors employed Bidirectional Encoder Representations from Transformers (BERT) to dy-

namically examine activity sequences in Android applications. They subsequently showed that BERT was particularly effective in handling imbalanced datasets.

Intrusion detection is among the Cybersecurity applications that are heavily affected by data imbalance. Hence, a remarkable number of research studies have been conducted recently towards this direction. Particularly, Bagui and Li examined how a variety of oversampling and undersampling techniques affect the performance of several multi-class neural network classifiers on a set of benchmark Cybersecurity datasets [11]. They found that these techniques increase Recall significantly only in cases of extreme data; otherwise, their impact is rather limited. Moreover, oversampling had a better performance on attack detection (minority data).

Azizjon et al. tackled class imbalance with a random oversampling technique and a one-dimensional Convolutional Neural Network (CNN) classifier [9]. In contrast, Fu et al. adopted the ADASYN oversampling method, introducing a sophisticated classifier called DLNID based on a CNN combined with Bi-Directional Long-Short Term Memory (BiLSTM) units [37].

Regarding the hybrid sampling methods, Zhang et al. utilized SMOTE for oversampling and a Gaussian Mixture Model to perform clustering on the samples of the majority class [118]. Another interesting hybrid approach was introduced by Ding et al. in [28], where the undersampling process was performed by a k-NN algorithm, whereas the oversampling was carried out by a Generative Adversarial Network (GAN) called TACGAN.

In fact, GANs have been proved quite effective in a wide variety of oversampling tasks. In the context of intrusion detection, the study of Lee and Park used GANs to perform oversampling of the minority class, and then, they used a Random Forest classifier to test the detection accuracy [63]. Other deep learning models have also been examined in the recent literature including the aforementioned CNNs [9, 118], CNN-BiLSTM hybrids [37], and Deep Neural Networks (DNN) with Variational Autoencoders (VAE) [3].

Fraud detection is another sub-area of Cybersecurity where class imbalance is observed. Makki et al. compared 8 machine learning methods, along with several re-sampling techniques to address class imbalance in credit card fraud classification [72]. Remarkably, the study concluded that the standard techniques for tackling data imbalance may have unpleasant consequences when the imbalance is extreme, because they generate a significant number of false positives. Finally, Fu et al. proposed a CNN-based fraud detection framework, to capture significant patterns of fraud behaviors learned from labeled data [36]. The authors utilized a massive transaction dataset to demonstrate the superiority of their model against standard machine learning classifiers like Random Forests, Neural Networks and SVMs.

3.2 Bioinformatics and Medical Sciences

The primary objective in Bioinformatics is to study and develop computational models with the aim of analyzing medical and biological data. When the underlying

data includes genomic information, the problem of class imbalance in classification tasks becomes too severe to be ignored. Examples of such tasks include disease diagnosis [100, 116, 58, 10], protein classification [123, 93], protein sub-cellular prediction [101], prediction of drug effectiveness, and so forth.

In one of the first studies in the area, Cohen et al. viewed the problem of hospital infections as a classification problem [26]. The 11:89 imbalance ratio of the underlying data motivated the authors to introduce, first, an effective hybrid re-sampling approach, and, second, a modified support vector classifier that was adjusted to improve the recognition of rare cases.

The work of Bugnon et al. studied multiple deep learning architectures in order to confront extreme data imbalance (ratios of up to 1:2000) in precursor microRNA classification [16]. The most interesting conclusion of this study is that, in presence of imbalanced data, the Deep Belief Networks (DBN) achieve superior performance than other deep learning models such as self-organizing maps (SOM). On a similar spirit, Bach et al. attested various data preprocessing techniques and classification algorithms to achieve decent performance in osteoporosis predictions [10]. They inferred that, for a dataset with an imbalance ratio of about 7:93, the highest accuracy was obtained by combining SMOTE with a Random Forest classifier. The Random Forest classifier has also been studied in an older paper authored by Dittman et al. [29]. In that work, the random undersampling technique was applied to augment 15 imbalanced bioinformatics datasets. In the context of Type 2 Diabetes Mellitus detection, Ramadhan compared SMOTE and ADASYN in conjunction with an SVM classifier and found ADASYN superior [83].

In [58], Krawczyk et al. presented a clinical decision support system for breast cancer malignancy grading. The system employs three segmentation algorithms to extract the most important features from clinical images, and feeds them to EUS-Boost, a novel ensemble classification model. The model itself brings balance to the underlying data via undersampling with boosting. Wan et al. introduced another ensemble classifier named HPSLPred to efficiently perform multi-label classification of protein sub-cellular localization [101]. The problem of data imbalance is tackled by applying an SVM-based algorithm that searches for balanced samples near the decision boundary. In [116], the authors utilized an asymmetric bagging (asBagging) ensemble classifier to classify high-dimensional imbalanced biomedicine data. Their model was improved by integrating an improved random subspace generation strategy that is called feature subspace (FSS).

More recently, Zhang et al. [121] introduced a new method called pseudo-negative sampling that is suitable for extremely imbalanced datasets. This work assumes that the negative samples constitute the majority class, and proposes a technique that converts negative (majority) samples into positive (minority) samples. These samples are called pseudo-negatives, and they are identified by applying a supervised method based on a max-relevance min-redundancy criterion beyond Pearson correlation coefficient. The method was experimentally attested on seven datasets by employing multiple classifiers, and it was found superior to SMOTE, max-relevance, and min-redundancy.

Finally, Song et al. devised a multi-stage method to effectively identify of DNA-binding proteins [93]. In the heart of this method, a feature extraction method generated 188-dimensional feature vectors based on the property of minimum Redundancy and Maximum Relevance to represent the protein structure. In the sequel, the vectors were fed into an ensemble classifier named imDC, and the final output was produced by a new predictor named nDNA-Prot. The performance of the proposed framework was experimentally attested against DNA-Prot and iDNA-Prot, and it was found superior in terms of accuracy and Area Under the Curve (AUC).

3.3 Natural Language Processing (NLP)

The extraction of useful knowledge from raw text and the capture of meanings in text collections have improved dramatically over the past few years. The tremendous growth of the deep learning research has been beneficial for Natural Language Processing. Nevertheless, the problem of data imbalance affects significantly the performance of the underlying models in this application area too.

In [80], Padurariu and Breaban presented an experimental analysis by using different classifiers, text representations and class balancing techniques to derive an effective model for classifying short job descriptions. The authors introduced a differential evolution algorithm to establish a cost-sensitive method. On the other hand, Yang et al. proposed a hybrid approach that modifies a standard CNN with the aim of handling the majority class, whereas it also applies few-shot learning techniques to handle the minority classes [112]. Castellanos et al. introduced a SMOTE-based iterative approach that applies to classification with imbalanced string data. More specifically, their method generates artificial strings that lie between two training samples without requiring the original data to be transformed into embeddings [19].

In an older work, Ogura et al. introduced a feature selection algorithm to confront data imbalance [79]. The authors utilized multiple metrics to determine the feature importance in text training data, and concluded that the signed versions of chi-squared and Information Gain are far superior to their unsigned counterparts. Another comparative study of a set of re-sampling methods in text classification with SVMs was conducted by Sun et al. [94]. Interestingly, the authors employed 10 such methods and found that none of them had a positive impact on the SVM effectiveness.

Nowadays, sentiment analysis is among the hottest NLP research topics. The polarity classification of user opinions in reviews, comments, blog posts, etc., is of crucial importance in many diverse applications [7]. To confront the distribution imbalance of review texts, Wang et al. introduced the BRC algorithm for under-sampling the majority class in the decision boundary region [104]. Kübler et al. evaluated several feature selection methods to alleviate the imbalance of user ratings on a collection of cooking recipes [60].

On a similar study, Obiedat et al. worked with customer reviews on restaurants. The authors tackle the problem of polarity imbalance by testing multiple oversam-

pling techniques, including SMOTE, two of its variants, and Adaptive Synthetic Sampling. The augmented data is eventually fed into a cost-sensitive SVM classifier with Particle Swarm Optimization [77]. The work of Ghosh et al. was focused on micro-blogging platforms [44]. The authors applied oversampling on the minority class samples, and they utilized SVM and Naive Bayes models for binary sentiment classification. Another oversampling approach based on the satisfaction of the distributional hypothesis was introduced by Moreo et al. [75].

It is apparent that the majority of works in the area either apply or enhance typical methods, such as re-sampling, cost-sensitive algorithm modifications, or hybrid approaches. In contrast, Abonizio et al. studied additional text augmentation methods such as back-translation and BART in conjunction with modern deep classification models (LSTMs, CNNs, GRUs, etc.) [5]. The experimental results revealed that data augmentation (particularly of the minority classes) can indeed lead to better sentiment classification performance. In addition, Liu et al. introduced a probabilistic term-weighting method to identify the documents of the minority classes effectively. More specifically, they devised two relevance indicators that were translated as probabilities reflecting the document class [70].

3.4 Computer Vision

This field studies AI algorithms that extract meaningful information from images and multimedia content like videos and streaming visual data. These algorithms are usually accompanied by augmentation techniques that pre-process the input data with the aim of enhancing their effectiveness.

The problem of imbalance in computer vision occurs when a large number of training images do not contain the objects of interest. Therefore, the performance of the utilized classification models is poor when these objects are present. To address this problem, Gao et al. introduced a step-wise hierarchical structure algorithm (called EHS) that performed sampling, filtering, and model training at each step. In comparison to the traditional oversampling and undersampling methods, EHS was proved to be superior in two TRECVID2010 datasets [42]. Yang and Chen proposed a new sampling technique that was subsequently combined with an ensemble learning process. The framework was named PEEL, and its effectiveness was experimentally demonstrated in the context of video event detection [113].

In Subsection 3.1 we have discussed about the effectiveness of GANs in addressing the issue of data imbalance in multiple applications. GANs have been utilized extensively with the aim of generating artificial samples for the minority classes. Hence, this model can also be viewed as an oversampling approach. In their recent study, Sampath et al. investigated the most recent advances in GAN-based research for imbalanced image data [87].

Regarding the popular problem of image classification, multiple research groups have focused on the introduction of methods that combat class imbalance. More specifically, Khan et al. proposed a cost-sensitive deep CNN to learn the most

important features from both the majority and minority classes [54]. The learning procedure of this architecture, called CoSen, jointly optimizes the class-dependent costs and the neural network parameters. In their study, Reza and Ma worked with two imbalanced medical image datasets about breast cancer [85]. The authors tested several oversampling and undersampling techniques with the aim of improving the accuracy of their CNN classifier, and highlighted the usefulness of oversampling.

The work of Huynh et al. [51] was also oriented towards medical imaging. However, it mainly applies to perturbation-based semi-supervised learning methods. ACBL is an adaptive model that includes a supervised part that optimizes the cross-entropy loss of a ResNet CNN and a semi-supervised part that minimizes the consistency loss between the class distributions of the unlabelled and the predicted image data. Before they are fed into the model, the unlabelled images are augmented by using the Unsupervised Data Augmentation (UDA) method.

In 2017, Odena et al. introduced ACGAN, a model based on Generative Adversarial Networks (GAN) with label conditioning [78]. ACGAN produces 128×128 resolution images that capture important class information. Nevertheless, this model was not specifically designed for imbalanced datasets. In their preprint, Mariani et al. proposed BAGAN, a model that, similarly to ACGAN, used label conditioning to perform augmentation of imbalanced image data [74]. BAGAN synthesizes images belonging to the minority class by using useful features that it learnt from its training with majority class images. The class distributions in the latent space are inferred by the encoder part of an autoencoder. This strategy allows the adversarial training to begin from a more stable point. The authors demonstrated the superior performance of BAGAN by comparing it with other state-of-the-art GANs including ACGAN.

In [62], the authors introduced a transfer learning approach for plankton image classification. At first, Random Oversampling was applied to the data to alleviate class imbalance. Then, a CNN was pre-trained on the balanced data, and the model was subsequently fine-tuned by using the original data. Pouyanfar et al. proposed a dynamic sampling approach that adapts itself according to how well a deep CNN has learnt a specific class [82]. In this way, the classes for which the accuracy is low are oversampled, whereas, in contrast, the high performance classes are undersampled.

Finally, Wardhani et al. [106] examined various evaluation metrics in the context of imbalanced image classification. Among a set of different choices, the authors conclude that Area Under the Curve (AUC) is the measure that most accurately reflects the performance of a classifier in the presence of imbalanced training data.

4 Technique-based Organization

In the previous section, we presented the most important research articles on data imbalance from an application-wise perspective. Of course, the literature also contains a significant number of works that examine the generic version of the problem. That is, without considering any particular application field. This section organizes both the generic and the application-specific studies on data imbalance according

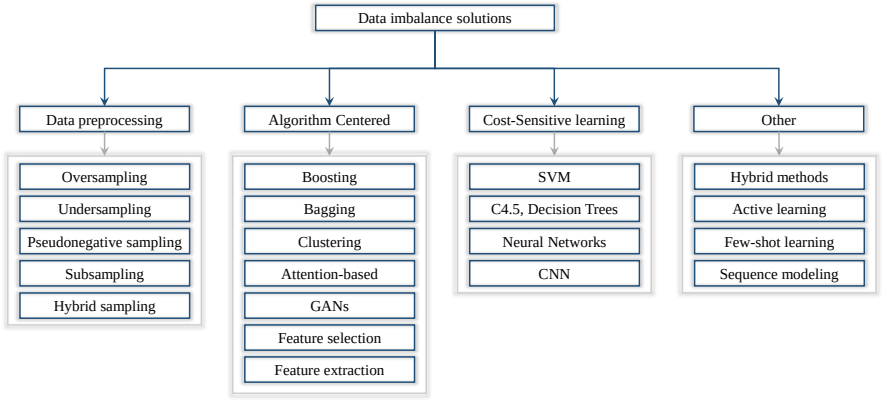

Fig. 3 Categorization of methods for handling data imbalance.

to the techniques that they introduce to address the problem. Since the latter have already been discussed, here we provide descriptions only for the generic papers.

The proposed methods are typically divided into four categories:

- *Data Preprocessing*: The methods of this category augment the imbalanced data either by appending artificial samples to the minority class (oversampling), or by removing samples from the majority class (undersampling). A third category includes hybrid sampling approaches that apply both oversampling and undersampling to render the data balanced. Moreover, augmentation can be achieved by applying either feature selection or feature extraction techniques, with the aim of identifying or generating meaningful features.
- *Algorithm Centered Approaches*: This family includes the algorithmic approaches to imbalanced data handling. Oversampling with Boosting or GANs and undersampling with clustering or feature selection/extraction are indicative representatives of such algorithms. We examine them together with the data preprocessing techniques in Subsection 4.1.
- *Cost-Sensitive Methods*: Here, we encounter techniques that directly modify a learning algorithm in order to decrease its bias towards the majority class. In particular, the cost-sensitive methods take into consideration the error of the misclassified samples. The literature includes methods for enhancing the most popular models including the SVMs, Neural Networks, Decision Trees, etc.
- *Hybrid Methods*: Several strategies adopt a hybrid approach that both modifies a classification algorithm and applies data preprocessing techniques to mitigate the effects of class imbalance. A portion of these methods may also apply novel learning strategies like active, or few-shot learning.

Figure 3 illustrates the aforementioned categorization. Table 2 will be our driver throughout this section. It displays a taxonomy of the relevant articles, according to the technique they introduce or apply to confront the problem of data imbalance. The horizontal lines are used to discriminate methods belonging to different categories.

Table 2 Data imbalance solutions categorized by application field.

Technique	Papers
Oversampling – Random	[3], [9], [11], [15], [30], [62], [75], [80], [85], [114], [120]
Oversampling – SMOTE	[6], [10], [11], [16], [18], [19], [22], [25], [35], [44], [64], [77], [80], [83], [85], [94], [109], [118], [123]
Oversampling – Borderline SMOTE	[24], [46], [67], [77], [91]
Oversampling – Safe-level SMOTE	[17], [27]
Oversampling – Adaptive SMOTE, Gaussian	[81]
Oversampling – SMOTE-SVM	[4], [77], [80]
Oversampling – MWMOTE	[14], [107], [108]
Oversampling – SNOCC	[124]
Oversampling – ADASYN	[6], [11], [37], [48], [77], [83], [85]
Oversampling – w/ Boosting	[23], [95]
Oversampling – w/ Bagging	[105]
Oversampling – w/ Attention	[102]
Oversampling – with GANs	[8], [28], [31], [63], [66], [74], [111]
Undersampling – Random	[3], [10], [11], [15], [29], [30], [42], [47], [85], [94], [109], [114]
Undersampling – ENN	[10], [110]
Undersampling – kNN	[28]
Undersampling - Wilson's Editing	[13], [88]
Undersampling – w/ Boosting	[40], [58], [95], [114]
Undersampling – w/ Clustering	[68], [115], [118]
Undersampling – w/ Feature Selection	[69], [70], [71], [79], [104]
Undersampling – w/ Feature Extraction	[93], [113], [116]
Subsampling – Random	[3]
Pseudo-negative sampling	[121]
Hybrid sampling	[11], [18], [26], [28], [33], [82], [88], [103], [110], [118]
Cost-sensitive Learning – C4.5, Decision Tree	[12], [25], [30], [34], [59], [69], [86], [98]
Cost-sensitive Learning – SVM	[25], [34], [56], [64], [77], [101]
Cost-sensitive Learning – MLP, DNN	[20], [34], [122], [125]
Cost-sensitive Learning – CNN	[38], [54], [92], [119]
Cost-sensitive Learning – Deep Belief Net	[117]
Few-shot Learning	[112]
Sequence Modeling – BERT	[76]
Hybrid – resampling + cost-sensitive	[26], [28], [42], [77]

4.1 Data Preprocessing and Algorithm Centered approaches

One of the first approaches to the problem dictates the augmentation of the initial data with the aim of limiting its intrinsic imbalance. The methods that belong to this category can be further organized into three fields: oversampling, undersampling, and hybrid sampling. Here, we choose to present the algorithm centered methods along with the pure data preprocessing techniques because, in most of them, the ultimate goal is to either enrich the minority class or compress the majority class.

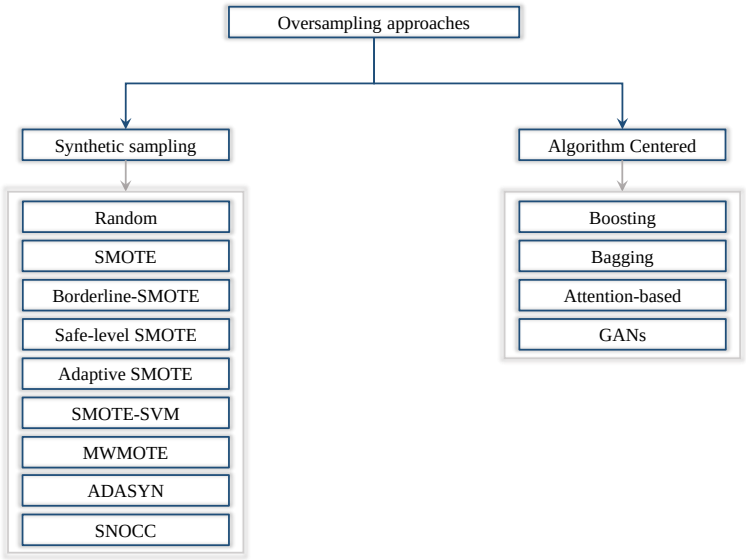

Fig. 4 Categorization of the oversampling techniques.

4.1.1 Oversampling

Oversampling refers to pre-processing strategies that enrich the minority classes with artificial samples, aiming at alleviating the problem of class imbalance. This simplistic definition includes a surprising number of oversampling approaches, such as typical data mining algorithms and heuristics, boosting, bagging, deep generative models, and so on. Figure 4 depicts a deeper categorization into two major groups: traditional synthetic sampling and algorithm centered techniques.

Random oversampling is perhaps the simplest data enrichment method: it repeatedly appends samples to the minority classes in a random fashion. The Random Walk Over-Sampling approach (RWO-Sampling) is a more sophisticated solution proposed by Zhang and Li [120]. RWO-Sampling is based on the proof that, under special circumstances, the expected average and standard deviation of the synthetic data can become equal to those of the original minority class samples. The method includes a mechanism that expands the minority class boundary after the creation of the synthetic samples. The authors have experimentally shown that RWO-Sampling significantly outperforms other methods when common classifiers are used.

In 2002 Chawla et al. introduced one of the most successful preprocessing methods, called Synthetic Minority Oversampling Technique (SMOTE) [22]. For each sample of the minority class, SMOTE initially identifies its k nearest neighbors (Chawla et al. suggested $k = 5$). Then, new synthetic samples are generated between each sample and its k nearest neighbors, until the dataset becomes balanced. In this way, SMOTE confronts the problem of overfitting and improves the generalization capabilities of a classifier.

SMOTE has been extensively studied and utilized in numerous researches involving imbalanced data [21, 43]. In addition, a huge number of extensions for it have been proposed in the literature. A detailed list of SMOTE extensions up to 2018 is provided in [35]. We will not reproduce this list here, but we will focus on the two most well-studied properties of SMOTE: i) the selection of the data points that will be oversampled, and ii) the creation of the synthetic samples (i.e., the interpolation type).

Regarding the initial samples selection, Han et al. introduced Borderline-SMOTE, a technique that creates synthetic samples only for those instances that lie close to the decision boundary between two classes [46]. In other words, Borderline-SMOTE does not operate on all the examples of the minority class, but it focuses only on those that are considered important for the output of a classifier. On the other hand, Safe-Level SMOTE computes a weight degree by taking into account the number of nearest neighbors of a minority sample [17]. This degree, called the safe level, determines the area where the synthetic samples will be created. Despite that more than a decade has passed since their introduction, both Borderline and Safe-level SMOTE are still being used to mitigate class imbalance in various applications [24, 27, 67, 91].

The interpolation type is the second property that is most studied in the literature. SMOTE and its two aforementioned variants randomly synthesize samples on the line that connects a minority sample with its nearest neighbor (linear interpolation). However, other approaches may create samples that are closer to an instance than its neighbors by employing feature weighting schemes [50], or topologies of different shapes (e.g., ellipses [2]).

One year after the introduction of SMOTE, Chawla et al. enhanced their technique with a Boosting mechanism called SMOTEBoost [23]. The method is a modification of the AdaBoost.M2 procedure for converting weak classifiers into strong ones. After SMOTE is applied to the original imbalanced data, an iterative process repeatedly trains a weak classifier by assigning higher weights to the misclassified samples.

Apart from boosting, bagging has also been employed to develop ensemble-based algorithms. For instance, Wang and Yao proposed SMOTEBagging, a SMOTE enhancement for effectively learning multi-class imbalanced datasets [105]. The review paper of Galar et al. studied multiple ensemble-based techniques, including boosting, bagging and hybrid methods [89].

In [14], the authors introduced a two-stage clustering-based approach called MWMOTE (Majority Weighted Minority Oversampling TEchnique). During the first stage, MWMOTE locates the nearest neighbours of the samples belonging to the minority class and assigns distance-based weights to them. In the sequel, it employs a clustering algorithm to generate synthetic data, under the restriction that the artificial data lie inside some minority class cluster. Wei et al. introduced a noise-resistant variant of the algorithm, called NI-MWMOTE [108]. Their method is based on an estimator that tries to predict whether a suspected noise is true by examining the neighborhood density. In the same year, the same group presented Cluster-MWMOTE that combined agglomerative clustering with MWMOTE [107].

Zheng et al. recognized two significant disadvantages in SMOTE: First, its linear interpolation property eventually produces data points that do not represent the original data distribution. And, second, simply searching for the k nearest neighbors of a minority sample may lead to the problem of overlapping between classes. For these reasons, they introduced an oversampling method called SNOCC (Sigma Nearest Oversampling based on Convex Combination) with the aim of curating the weaknesses of SMOTE [124]. In SNOCC, the samples to be created are determined by a convex combination function of its nearest neighbors. The authors demonstrate that their approach can reproduce the distribution of original data more accurately, even if the distribution is irregular.

Another family of approaches applied SMOTE only to the support vectors of a class. Initially, an SVM classifier is trained on the original data. Then, SMOTE creates synthetic samples by taking into consideration only the computed support vectors. This method, called SMOTE-SVM, cannot balance the dataset completely. However, it appropriately adjusts the decision boundary between two classes [4, 80].

In [81], Pan et al. introduced two oversampling techniques, Adaptive SMOTE and Gaussian oversampling. The former synthesizes a new minority class by adaptively selecting instances from the original minority class, with the aim of enhancing the distributional data characteristics. The latter, establishes a Gaussian distribution by sampling the strong characteristics of the majority class and by applying dimensionality reduction on the original imbalanced data. The authors evaluated both methods by using 15 datasets and proved their usefulness against other traditional techniques.

He et al. introduced a novel oversampling method, named ADASYN (Adaptive Synthetic Sampling), that produces more synthetic examples for the input samples that are harder to learn [48]. To achieve this goal, ADASYN assigns different weights to the samples by enumerating their k nearest neighbors that belong to the majority class. In this way, ADASYN adaptively moves the decision boundary towards the difficult examples, thus decreasing the classification bias.

The great capability of the recent state-of-the-art generative models to synthesize samples that represent the probability distribution of the original data, rendered them particularly attractive for oversampling purposes. As already mentioned, the Generative Adversarial Networks (GANs) have been effectively applied in intrusion detection systems [28, 63] and image classification applications [74, 78]. In [8], the authors introduced MFC-GAN, a model that integrates not one, but multiple fake classes to ensure better data generation. Yan and Zhou initialized the training process of their IDA-GAN model by exploiting a Variational Autoencoder to learn the majority and the minority class distributions [111]. Wei et al. proposed EID-GAN as a remedy to the BAGAN's weakness in synthesizing tiny outliers [66]. Finally, Engelmann and Lessmann used a conditional Wasserstein GAN with the aim of modeling tabular datasets with numerical and categorical variables [31].

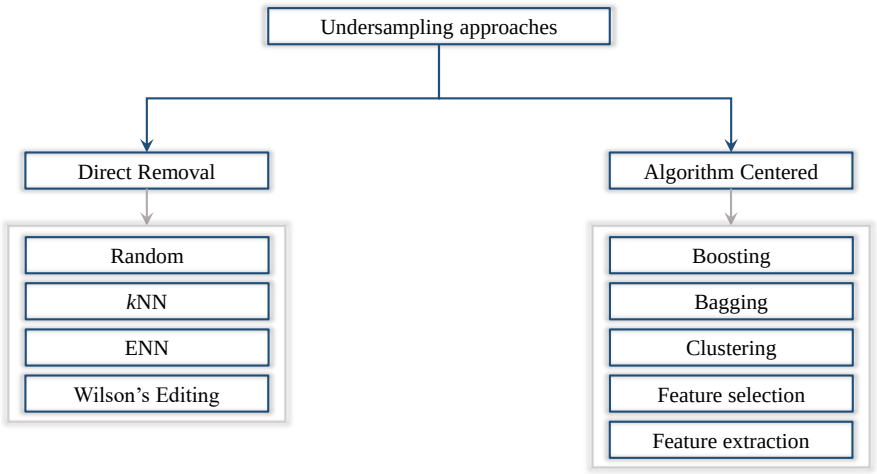

Fig. 5 Categorization of the undersampling techniques.

4.1.2 Undersampling

Undersampling is the preprocessing technique that brings balance to a dataset via sample removal from the majority class. On an early work, Drumond and Holte compared the impact of undersampling and oversampling on the performance of a C4.5 decision tree classifier [30]. Their analysis indicated that oversampling is inferior, because the performance discrepancies are usually infinitesimal when the misclassification costs are changed. However, note that the undersampling techniques run the risk of breaking the majority class distribution by discarding important samples from it.

Similarly to the oversampling case, random undersampling removes instances from the majority class in a random fashion. Hasanin and Khoshgoftaar studied the effects of random undersampling on classification performance by using five large-scale datasets and the Random Forest classifier of the Apache Spark framework [47]. They inferred that performance can be enhanced by randomly increasing the minority class percentages from 0.1% to 1.0% and partially undersampling the majority class, without balancing the data to a 50:50 ratio.

Wilson's Editing is a simple method introduced by Barandela et al. in 2004 [13]. It is based on a lazy kNN classifier, with $k = 3$, that prunes all the misclassified majority samples. On the other hand, RUSBoost is one of the most popular undersampling techniques [89]. It combines random undersampling with AdaBoost to mitigate class imbalance, and it often performs better than SMOTE and SMOTEBoost. It is also a simpler and faster method. Galar et al. also capitalized on Boosting and proposed EUSBoost, an evolutionary undersampling approach that was subsequently combined with AdaBoost.M2 applied at C4.5 classifiers [40].

Clustering is another effective way for undersampling the majority class on imbalanced datasets. The idea is that these algorithms construct homogeneous, complete clusters that group together similar data points. Therefore, an entire cluster of elements can be effectively replaced by a single representative element, e.g., cluster center, centroid, clustroid, etc. This key concept renders clustering an attractive approach to undersampling, because it enables the replacement of entire sample groups by a single, representative data point.

Lin et al. proposed two clustering strategies for undersampling, both based on the well-known k-Means algorithm [68]. In the first strategy, the number of clusters to be constructed is determined by the population of the minority samples. The centroids of the generated clusters are subsequently utilized to replace the entire majority class, rendering the dataset perfectly balanced. The second strategy is very similar to the first one, except that we do not use the centroids as representative points, but their nearest neighbor. To demonstrate the effectiveness of their methods, the authors employed 44 small and 2 large datasets to train 5 different classifiers.

In contrast, the work of Lin et al., the clustering approaches of Yen and Lee are applied to the entire dataset, leading to mixed clusters that contain samples from both the minority and majority classes [115]. In the sequel, the representative majority class samples are selected by adopting a variety of different criteria, e.g., the most proximal or the farthest points to the minority class samples.

Several researchers have introduced methods that combine undersampling with feature selection techniques. These techniques have been proved valuable in many studies because a careful selection of features can improve the accuracy of the minority class prediction.

More specifically, the authors of [71] introduced an ensemble learner that combines evolutionary under-sampling with feature selection. The selection is performed by using the multiobjective ant colony optimization algorithm that maximizes F1 and G-Mean. Maldonado et al. investigated feature selection from the perspective of dimensionality reduction in imbalanced datasets [73]. They introduced a feature elimination technique by repeatedly training an SVM classifier on a feature set that was progressively refined. The non-important input variables were removed from the feature set on each iteration. However, to improve the performance of the SVM predictor, the algorithm performed oversampling on the minority class samples by using SMOTE.

The feature selection method of Liu et al. utilized a Decision Tree classifier and used its splitting criterion to identify the most informative features [69]. The authors proposed a Weighted Gini index, with the aim of increasing the bias towards the minority class. Consequently, the method of [69] can be also considered as a cost-sensitive approach, and that explains why this paper appears twice in Table 2.

4.1.3 Hybrid Sampling

In the two previous subsections, we mentioned that, in several cases, oversampling has infinitesimal impact on the performance of a classifier, whereas it can also

distort the probability distribution of the minority classes. On the other hand, the undersampling methods can be detrimental, as they may discard important samples from the majority class. To overcome these issues, several scholars proposed the usage of hybrid sampling techniques to mitigate class imbalance.

The method of Cao and Zhai was based on a combination of SMOTE (for oversampling) and random undersampling [18]. The experimental evaluation, with a binary SVM classifier trained on five datasets, demonstrated the superiority of this simple method over the cases that employ solely either random underampling or SMOTE. In [110], the authors presented the RFMSE hybrid sampling algorithm that combines Misclassification-Oriented SMOTE and the Edited Nearest Neighbor (ENN) method for noise removal. RFMSE was tested on 10 UCI imbalanced datasets and outperformed 9 sampling algorithms in 9 cases.

Zhu et al. [126] based their work on the concept that the instances that are close to the decision boundary (overlapping region) play a greater role in classification effectiveness than the more distant ones. They presented an Evolutionary Hybrid Sampling technique (EHSO) that attempts to "strengthen" the decision boundary by dropping the non-informative samples of the majority class. The CHC genetic algorithm is applied for this reason [32], whereas random oversampling is applied to synthesize new minority samples.

The hybrid method of [103] initially trained an SVM classifier by using the original imbalanced dataset. Then, it performed undersampling, by removing majority class samples according to their distance from the decision boundary determined by the SVM classifier. The most distant samples were removed, until the imbalance ratio became the half of the one of the original dataset. In the sequel, the new dataset was divided into k parts, and one of them was randomly chosen. SMOTE was applied to it, and a new SVM classifier was trained on that balanced dataset. This process was repeated for the rest $k - 1$ splits.

In [88], Seiffert et al. examined three hybrid resampling scenarios on 10 datasets from different application fields. The first scenario initially applied oversampling to the minority class by using SMOTE or Borderline SMOTE and, then, it performed random undersampling. The second use case reversed the aforementioned process, whereas the third procedure applied SMOTE, Borderline SMOTE, random undersampling, and random oversampling, in combination with Wilson's Editing. The experiments verified that hybrid sampling is superior to the individual sampling techniques. It never performed remarkably worse than the individual methods, and, in the vast majority of cases, it improved the performance of a C4.5 Decision Tree predictor.

Estabrooks et al. presented a comparative study of oversampling and undersampling by utilizing the well-established Reuters-21578 text collection [33]. It was demonstrated that, by combining various resampling techniques, better results can be obtained.

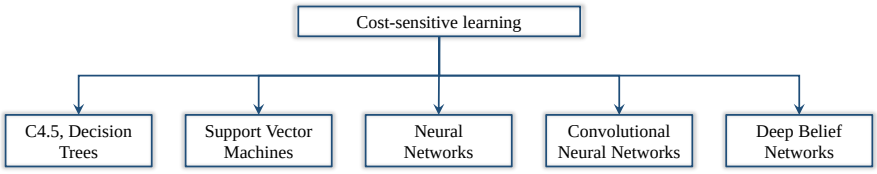

Fig. 6 Categorization of the cost-sensitive techniques.

4.2 Cost-Sensitive Learning

The typical machine learning models adopt a 0-1 cost assignment approach, that outputs 0 for correct predictions and 1 for the incorrect ones. This strategy works well when the classes are relatively balanced, but it becomes ineffective when the class distributions are skewed. The cost-sensitive methods introduce a special misclassification error that strongly penalizes a classifier for mispredicting the minority class samples. For this reason, they utilize a misclassification matrix C, whose entries C_{ij} represent the cost of assigning the j-th class to a sample x that originally belongs to the i-th class [34]. The definition of the contents of C is among the greatest challenges posed by these methods. In some applications, the individual misclassification costs are manually provided by experts.

The SVMs were among the first models that were enhanced with cost-sensitive mechanisms. We have already cited the recent works of Obiedat et al. [77] and Ghosh et al. [44] on sentiment analysis applications. Kim and Sohn introduced NN-CSSVM, a hybrid model comprised of neural networks and cost-sensitive SVMs [56]. The authors modified two real-world datasets to create five synthetic ones, by setting different imbalance ratios. NN-CSSVM outperformed the standalone cost-sensitive SVM and cost-sensitive Multilayer Perceptron by a significant margin. A cost sensitive-SVM was also introduced in [64].

Krawczyk et al. introduced an ensemble model, with Decision Trees as base learners [59]. Initially, the proposed technique randomly divides the original feature space into subspaces and, then, it trains each individual learner on a different feature subspace. This strategy improves the diversity of each Decision Tree. An evolutionary algorithm is subsequently executed to select the base learners by assigning weights to them. The model was tested on 6 benchmark datasets, and exhibited superior performance compared to SMOTEBagging, SMOTEBoosting and other methods. One year earlier, Sahin et al. [86] proposed a similar cost-sensitive Decision Tree for credit card fraud detection applications.

The work of Bahnsen et al. introduced an example-dependent cost-sensitive decision tree that takes into consideration different costs during node splitting [12]. More specifically, a misclassification cost matrix C with 4 elements (True/False-Positive/Negative cost) is initially employed. Then, the authors propose: i) a cost-sensitive impurity measure to determine the optimal splitting of a leaf, and ii) a cost-sensitive pruning policy that removes the tree nodes that do not contribute to

the minimization of the cost. In an older work, Ting proposed a cost-sensitive method that assigned weights to the input instances during Decision Tree induction [98].

In the relevant literature, we also encounter researchers who proposed cost-sensitive modifications of neural networks. In a earlier work (2002), Zhou and Liu presented an empirical study on how oversampling, undersampling and threshold moving affect the training process of cost-sensitive neural networks [125]. In their experimental evaluation, the authors employed three misclassification matrices according to the aforementioned work of Ting [98]. Despite its age, the conclusions of this paper are noteworthy: i) cost-sensitive learning is considerably more difficult on multiclass tasks than on two-class tasks, ii) SMOTE and several other resampling techniques alone are not effective in multiclass tasks, and iii) the ensemble methods (bagging and boosting) that employ resampling perform significantly better.

Zhao et al. introduced SPFCNN, a cost-sensitive model based on neural networks, aiming at detecting software defects [122]. The proposed model comprised two Siamese fully-connected networks, one shallow and one deep, trained with AdamW. The cost-sensitivity was integrated to the model through the normalized expected cost of misclassification. The experiments conducted on six test datasets demonstrated the effectiveness of SPFCNN.

The excellent performance of CNNs in the domain of image classification motivated several research groups to introduce cost-sensitive modifications of CNNs, with the goal of mitigating class imbalance in image datasets. The CoSen model of Khan et al. has been already described in Subsection 3.4 [54]. Fuqua and Razzaghi introduced another cost-sensitive CNN variant, named CSCNN, to address the problem of control chart pattern recognition [38]. CSCNN is based on the cost-sensitive function of [61] that penalizes differently the mispredictions of the minority and majority classes. Similarly to the well-known LeNet 5 model, the architecture includes sequences of 1D convolutional layers, followed by max pooling and dropout layers. The authors chose the mean squared propagation (RMSProp) algorithm for training CSCNN due to its robustness. A similar architecture, also called CSCNN, was presented in [92].

In [119], Zhang et al. introduced a cost-sensitive Residual CNN, named CS-ResNet, for the effective detection of cosmetic defects. Similarly to other CNNs, CS-ResNet contains typical convolutional layers with batch normalization units and max pooling. However, it also embodies a cost-sensitive adjustment layer for assigning class-dependent misclassification costs, based on a weighted softmax cross-entropy loss function.

The authors of [117] recognized that the conventional Deep Belief Networks (DBN) do not perform well on classification tasks that involve imbalanced data. So, they introduced an evolutionary cost-sensitive DBN (ECS-DBN) that initially applies an adaptive differential evolution algorithm to optimize the misclassification costs. In the sequel, the optimized costs are used to the DBN. The conducted experiments demonstrated that ECS-DBN is superior to resampling (SMOTE, Borderline SMOTE and SMOTE-SVM) and using a typical DBN for classification.

Table 3 Data imbalance solutions categorized by the utilized classification models.

Classifier	Papers
*k*NN	[14], [19], [52], [68], [79], [99], [108]
Naive Bayes	[3], [22], [44], [52], [68], [70], [97], [99], [126]
Logistic Regression	[80], [99], [109]
Support Vector Machines (SVM)	[5], [16], [18], [25], [42], [44], [52], [56], [64], [68], [70], [73], [77], [80], [83], [94], [99], [104], [107], [108]
Decision Trees, C4.5	[12], [14], [22], [25], [30], [33], [40], [52], [59], [68], [69], [80], [86], [88], [98], [99], [109], [126]
Random Forests	[3], [5], [10], [29], [47], [52], [63], [76], [99], [108], [109], [110]
Multilayer Perceptrons (MLP), Deep Neural Networks (DNN)	[3], [6], [11], [14], [16], [20], [28], [56], [68], [99], [102], [108], [109], [115], [122], [125]
Convolutional Neural Networks (CNN)	[5], [8], [9], [15], [36], [37], [38], [54], [62], [82], [85], [92], [112], [118], [119]
ResNet-18	[74]
Long Short-Term Memory (LSTM), Bi-directional LSTM (BiLSTM)	[5], [9], [76]
CNN + LSTM Hybrid	[9]
DLNID (CNN + BiLSTM Hybrid with attention)	[37]
Gated Recurrent Unit (GRU)	[5]
Deep Belief Networks (DBN)	[16], [117]
Deep Self-Organizing Maps (SOM)	[16]
Transformers	[5]
Ensemble (Bagging, Boosting, Hybrid)	[40], [58], [64], [68], [71], [89], [93], [101], [105], [113], [114], [116], [123]

5 Classifier-based Organization

As mentioned earlier, the a large part of the research works that are dealing with the problem of data imbalance primarily focus on classification tasks. More specifically, these articles either employ or adapt a variety of classification models to examine how the problem in question affects their performance. To cover this aspect, this section presents the most important works in the area from the perspective of the utilized classifiers.

Table 3 summarizes our classifier-based organization. The majority of the included articles have already been described in previous sections, so we shall not discuss them again here. Another portion of them represents survey and review papers, and their details will be presented in Section 6.

A quick observation of the Table 3 immediately reveals that the relevant research has been heavily based on traditional machine learning predictors, like Decision Trees, Random Forests, Naive Bayes, and Support Vector Machines. Despite their excellent performance and high popularity across numerous diverse application fields, the deep learning models have not been used extensively in data imbalance papers. The Convolutional Neural Networks (CNNs) and several deep CNN-based

architectures are perhaps the only exception to this observation. This is a strong conclusion, that is also verified by the survey of Johnson & Khoshgoftaar [52] on deep learning methods. It highlights the necessity of integrating such models in the techniques that confront data imbalance.

6 Surveys and Overviews

The vast number of research works that deal with the problem of classification with imbalanced data established the necessity for conducting comparative surveys in order to cover the most recent advances in the area. During our research, we discovered at least 22 such qualitative overviews, with the oldest being 16 years old. Table 4 contains a comparative summary of these surveys, ordered in ascending chronological order. The third column provides an impression of each work's length, whereas the last column contains descriptive snippets. The table itself reveals that, since 2007, there is at least one major survey paper about the research field of data imbalance.

All survey papers describe the current state-of-the-art solutions, the most significant challenges of the problem and the involved applications. Therefore, we shall not describe all of them here, because, more or less, they all approach the problem from a similar point of view. In contrast, we do focus on experimental surveys. That is, works that, apart from algorithm descriptions, also provide useful conclusions based on real-world applications and experiments.

The experimental survey of Galar et al. focused on the evaluation of multiple bagging and boosting algorithms on 44 binary-class imbalanced datasets [89]. In Section D, a very informative discussion of the conclusions is provided. In summary, the survey inferred that: i) SMOTEBagging, RUSBoost, and UnderBagging achieved the best performance among other ensemble learning methods, ii) the complex bagging and boosting methods do not lead to better results than the simpler ones, and iii) the bagging techniques are powerful, albeit difficult to implement. Another comparative study on the resampling techniques was conducted on 2007 by Van Hulse et al. [99]. The authors employed 35 datasets and several traditional machine learning models (see Table 3). The experiments revealed multiple high-performing combinations, such as Random Undersampling with C4.5 Decision Trees and Random Forests, and Random Oversampling with Logistic Regression.

The paper of Yap et al. investigated how oversampling, undersampling, bagging and boosting can improve classification performance on a cardiac surgery dataset with 4976 samples and an imbalance ratio of about 4:96 [114]. By using a Decision Tree classifier, the authors inferred that the resampling techniques worked significantly better than bagging and boosting. On the other hand, the work of Khoshgoftaar et al. studied how Random Forests (RF) can be fine-tuned to classify imbalanced data [55]. The experiments were conducted on 10 datasets of different sizes, imbalance ratios, and dimensionality. The authors used various values for two key RF attributes, the number of estimators in the forest and the number of features to be used during

Table 4 A comparative summary of overview, survey, and review papers on data imbalance.

Survey	Year	Number of papers	Summary
[99]	2007	18	Experimental; sampling methods and classifiers.
[55]	2007	24	Experimental; a Random Forest was fined-tuned on 10 datasets.
[96]	2009	92	Application fields; classification challenges; state-of-the-art solutions; evaluation measures.
[49]	2009	145	Application fields; classification challenges; state-of-the-art solutions; evaluation measures.
[94]	2009	37	Experimental; 10 sampling methods were evaluated with an SVM classifier on 3 text datasets; the results were negative.
[21]	2010	70	Sampling techniques and evaluation measures.
[39]	2011	112	Experimental survey on bagging, boosting, and hybrid state-of-the-art algorithms.
[41]	2012	30	Brief overview of several data preprocessing techniques and algorithm-based methods.
[1]	2013	67	State-of-the-art solutions.
[114]	2013	23	Experimental evaluation of oversampling, undersampling, bagging and boosting on a cardiac surgery dataset.
[84]	2014	57	Classification challenges; state-of-the-art solutions; evaluation measures.
[57]	2016	69	Application fields and challenges of learning from imbalanced data.
[45]	2017	337	Application fields; classification challenges; state-of-the-art solutions; Rare event detection.
[10]	2017	74	Experimental study on numerous oversampling and undersampling techniques in the field of osteoporosis prediction.
[90]	2017	13	Brief descriptions of several sampling techniques.
[35]	2018	238	This paper marked the 15-year anniversary of SMOTE. It discussed numerous SMOTE extensions and variants, as well as future insights for using SMOTE in Big Data applications.
[65]	2018	77	State-of-the-art solutions between 2010 and 2018 for extremely imbalanced datasets (ratios ranging from 100:1 to 10000:1).
[15]	2018	65	Experimental; exploration of oversampling, undersampling, two-phase training, and thresholding in image classification tasks.
[52]	2019	131	Deep learning models with imbalanced data; implementation details; experimental results.
[53]	2019	168	Application fields; classification challenges; state-of-the-art solutions; evaluation measures.
[97]	2020	48	Experimental; the impact of class imbalance on classification performance.
[87]	2021	236	Oversampling with Generative Adversarial Networks on computer vision applications.

training. They inferred that the best settings were 100 and $\lfloor \log_2 M + 1 \rfloor$, respectively, where M is the actual number of features in each dataset.

The survey of Leevy et al. summarized the state-of-the-art works on extreme data imbalance between 2010 and 2018 [65]. The work focuses on large-scale datasets that exhibit high majority-to-minority class ratios, between 100:1 and 10000:1. The large-scale study of Haixiang et al. explored 527 research articles on learning from imbalanced data, from the perspective of rare event detection [45]. The authors

introduced a generic data mining model that includes techniques for preprocessing the imbalanced data, classification and evaluation. Thabthah et al. have recently explored the relationship between the imbalance degree and the performance of a Naive Bayes classifier [97].

The brief overview of Ganganwar noticed the transition of data imbalance research toward the hybrid methods [41], whereas Shelke et al. discussed several oversampling and undersampling techniques [90]. Moreover, Krawczyk summarized the most crucial challenges of learning from imbalanced data and provided informative insights about the future solutions [57]. His work was not limited to classification applications only, but it also covered additional machine learning problems, including regression, clustering and big data analytics.

Buda et al. examined how oversampling, undersampling and two-phase training affect the performance of a CNN on image classification tasks [15]. By using three benchmark datasets, namely, MNIST, CIFAR-10 and ImageNet, the authors verified that data imbalance is indeed harmful, and inferred that oversampling: i) was the best-performing method in all cases, ii) must be fully applied until the dataset becomes perfectly balanced, and iii) does not cause overfitting to CNNs.

Johnson and Khoshgoftaar published a systematic survey on the research works that trained deep learning predictors with imbalanced datasets [52]. Interestingly, this work reports implementation details and experimental results for the most significant articles that it covers, providing a convenient way of quickly summarizing the key findings. The investigation concludes that, despite the huge adoption of the deep learning models, very little work has been done in the context of data imbalance. Even worse, the characteristics of Big Data are scarcely taken into consideration.

The recent large-scale study of Sampath et al. investigated the current state-of-the-art methods for oversampling the minority classes in computer vision tasks [87]. Motivated by the great capability of the Generative Adversarial Networks (GANs) to generate samples that reflect the original data distribution, the survey focused on oversampling methods that solely use GANs to achieve their goal. Due to their adversarial training, GANs have been proved particularly successful in generating synthetic images, and, thus, bringing balance to imbalanced datasets. Specifically, the study in question categorizes the relevant works according to the imbalance type they tackle, namely image, object, and pixel imbalance.

7 Conclusions

In this chapter, we presented on overview of the current state-of-the-art techniques for mitigating the problem of data imbalance in classification tasks. We adopted a triadic categorization approach for the relevant papers, namely, application-wise, technique-wise, and classifier-wise. Moreover, a summary of the most qualitative theoretical and experimental surveys was presented. We particularly emphasized on the conclusions of each work to assist the research community in quickly deriving conclusions from the involved methods.

Among the dozens of conclusion that we inferred from this overview, we mainly distinguish three key points:

1. Despite their proved effectiveness in multiple research fields, the deep learning models have not been used extensively in the problem of data imbalance yet. This inference is largely supported by our classifier-based analysis (Table 3) and the survey of Johnson & Khoshgoftaar [52]. Isolated exceptions include the usage/modification of Convolutional Neural Nets in image classification tasks, and several deep learning models in intrusion detection systems. Nonetheless, much work is still required to effectively integrate the powerful features of deep learning to methods that confront data imbalance.
2. The Generative Adversarial Networks (GANs) are quickly gaining the attention of the research community in oversampling tasks. Multiple such models have been presented in the literature to augment the minority classes with useful samples. Many more researchers are still trying to devise novel GAN-based methods for this purpose.
3. Complex solutions do not necessarily yield better results compared to their simpler (and older) counterparts. A large number of recent works are still employing baseline solutions, like SMOTE and its variants, or random undersampling and its variants, to restore balance in imbalanced datasets.

References

1. Abd Elrahman, S.M., Abraham, A.: A review of class imbalance problem. Journal of Network and Innovative Computing **1**(2013), 332–340
2. Abdi, L., Hashemi, S.: To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE Transactions on Knowledge and Data Engineering **28**(1), 238–251 (2015)
3. Abdulhammed, R., Faezipour, M., Abuzneid, A., AbuMallouh, A.: Deep and machine learning approaches for anomaly-based intrusion detection of imbalanced network traffic. IEEE Sensors Letters **3**(1), 1–4 (2018)
4. Abidine, M.B., Fergani, B.: Comparing HMM, LDA, SVM and Smote-SVM algorithms in classifying human activities. In: Proceedings of the 2015 Mediterranean Conference on Information & Communication Technologies, pp. 639–644 (2016)
5. Abonizio, H.Q., Paraiso, E.C., Barbon, S.: Toward text data augmentation for sentiment analysis. IEEE Transactions on Artificial Intelligence **3**(5), 657–668 (2021)
6. Akash, B.S., Yannam, P.K.R., Ruthvik, B.V.S., Kumar, L., Murthy, L.B., Krishna, A.: Predicting Cyber-Attacks on IoT Networks Using Deep-Learning and Different Variants of SMOTE. In: Proceedings of the 36th International Conference on Advanced Information Networking and Applications, pp. 243–255 (2022)
7. Akritidis, L., Bozanis, P.: Improving opinionated blog retrieval effectiveness with quality measures and temporal features. World Wide Web **17**(4), 777–798 (2014)
8. Ali-Gombe, A., Elyan, E.: MFC-GAN: Class-imbalanced dataset classification using multiple fake class generative adversarial network. Neurocomputing **361**, 212–221 (2019)
9. Azizjon, M., Jumabek, A., Kim, W.: 1D CNN based network intrusion detection with normalization on imbalanced data. In: Proceedings of the 2020 International Conference on Artificial Intelligence in Information and Communication, pp. 218–224 (2020)
10. Bach, M., Werner, A., Żywiec, J., Pluskiewicz, W.: The study of under-and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. Information Sciences **384**, 174–190 (2017)

11. Bagui, S., Li, K.: Resampling imbalanced data for network intrusion detection datasets. Journal of Big Data **8**(1), 1–41 (2021)
12. Bahnsen, A.C., Aouada, D., Ottersten, B.: Example-dependent cost-sensitive decision trees. Expert Systems with Applications **42**(19), 6609–6619 (2015)
13. Barandela, R., Valdovinos, R.M., Sánchez, J.S., Ferri, F.J.: The imbalanced training sample problem: Under or over sampling? In: Proceedings of the 2004 Joint IAPR International Workshops: Structural, Syntactic, and Statistical Pattern Recognition, pp. 806–814 (2004)
14. 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)
15. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural networks **106**, 249–259 (2018)
16. Bugnon, L.A., Yones, C., Milone, D.H., Stegmayer, G.: Deep neural architectures for highly imbalanced data in Bioinformatics. IEEE Transactions on Neural Networks and Learning Systems **31**(8), 2857–2867 (2019)
17. Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C.: Safe-Level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, pp. 475–482 (2009)
18. Cao, L., Zhai, Y.: Imbalanced data classification based on a hybrid resampling SVM method. In: Proceedings of the 12th IEEE International Conference on Ubiquitous Intelligence and Computing, and 12th IEEE International Conference Autonomic and Trusted Computing, and 15th IEEE International Conference on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), pp. 1533–1536 (2015)
19. Castellanos, F.J., Valero-Mas, J.J., Calvo-Zaragoza, J., Rico-Juan, J.R.: Oversampling imbalanced data in the string space. Pattern Recognition Letters **103**, 32–38 (2018)
20. Castro, C.L., Braga, A.P.: Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data. IEEE Transactions on Neural Networks and Learning Systems **24**(6), 888–899 (2013)
21. Chawla, N.V.: Data mining for imbalanced datasets: An overview. Data Mining and Knowledge Discovery Handbook pp. 875–886 (2010)
22. 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)
23. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: SMOTEBoost: Improving prediction of the minority class in boosting. In: Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 107–119 (2003)
24. Chen, Y., Chang, R., Guo, J.: Effects of data augmentation method borderline-SMOTE on emotion recognition of EEG signals based on convolutional neural network. IEEE Access **9**, 47491–47502 (2021)
25. Chen, Z., Yan, Q., Han, H., Wang, S., Peng, L., Wang, L., Yang, B.: Machine learning based mobile malware detection using highly imbalanced network traffic. Information Sciences **433**, 346–364 (2018)
26. Cohen, G., Hilario, M., Sax, H., Hugonnet, S., Geissbuhler, A.: Learning from imbalanced data in surveillance of nosocomial infection. Artificial Intelligence in Medicine **37**(1), 7–18 (2006)
27. Daud, S.N.S.S., Sudirman, R., Shing, T.W.: Safe-level SMOTE method for handling the class imbalanced problem in electroencephalography dataset of adult anxious state. Biomedical Signal Processing and Control **83**, 104649 (2023)
28. Ding, H., Chen, L., Dong, L., Fu, Z., Cui, X.: Imbalanced data classification: A KNN and Generative Adversarial Networks-based hybrid approach for intrusion detection. Future Generation Computer Systems **131**, 240–254 (2022)
29. Dittman, D.J., Khoshgoftaar, T.M., Napolitano, A.: The effect of data sampling when using random forest on imbalanced bioinformatics data. In: Proceedings of the 2015 IEEE International Conference on Information Reuse and Integration, pp. 457–463 (2015)

30. Drummond, C., Holte, R.C.: C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling. In: Proceedings of the 2003 Workshop on Learning from Imbalanced Datasets II, International Conference on Machine Learning, vol. 11, pp. 1–8 (2003)
31. Engelmann, J., Lessmann, S.: Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. Expert Systems with Applications **174**, 114582 (2021)
32. Eshelman, L.J.: The CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination. In: Foundations of genetic algorithms, vol. 1, pp. 265–283 (1991)
33. Estabrooks, A., Jo, T., Japkowicz, N.: A multiple resampling method for learning from imbalanced data sets. Computational Intelligence **20**(1), 18–36 (2004)
34. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from imbalanced data sets, vol. 10. Springer (2018)
35. Fernández, A., Garcia, S., Herrera, F., Chawla, N.V.: SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research **61**, 863–905 (2018)
36. Fu, K., Cheng, D., Tu, Y., Zhang, L.: Credit card fraud detection using Convolutional Neural Networks. In: Proceedings of the 23rd International Conference on Neural Information Processing, pp. 483–490 (2016)
37. Fu, Y., Du, Y., Cao, Z., Li, Q., Xiang, W.: A deep learning model for network intrusion detection with imbalanced data. Electronics **11**(6), 898 (2022)
38. Fuqua, D., Razzaghi, T.: A cost-sensitive Convolution Neural Network learning for control chart pattern recognition. Expert Systems with Applications **150**, 113275 (2020)
39. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **42**(4), 463–484 (2011)
40. Galar, M., Fernández, A., Barrenechea, E., Herrera, F.: EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. Pattern Recognition **46**(12), 3460–3471 (2013)
41. Ganganwar, V.: An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering **2**(4), 42–47 (2012)
42. Gao, Z., Zhang, L.f., Chen, M.y., Hauptmann, A., Zhang, H., Cai, A.N.: Enhanced and hierarchical structure algorithm for data imbalance problem in semantic extraction under massive video dataset. Multimedia Tools and Applications **68**, 641–657 (2014)
43. García, S., Luengo, J., Herrera, F.: Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowledge-Based Systems **98**, 1–29 (2016)
44. Ghosh, K., Banerjee, A., Chatterjee, S., Sen, S.: Imbalanced Twitter sentiment analysis using minority oversampling. In: Proceedings of the 10th IEEE International Conference on Awareness Science and Technology, pp. 1–5 (2019)
45. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications **73**, 220–239 (2017)
46. Han, H., Wang, W.Y., Mao, B.H.: Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In: Proceedings of 2005 International Conference on Intelligent Computing (Advances in Intelligent Computing), pp. 878–887 (2005)
47. Hasanin, T., Khoshgoftaar, T.: The effects of random undersampling with simulated class imbalance for big data. In: Proceedings of the 2018 IEEE International Conference on Information Reuse and Integration, pp. 70–79 (2018)
48. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive Synthetic Sampling approach for imbalanced learning. In: Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328 (2008)
49. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering **21**(9), 1263–1284 (2009)

50. Hukerikar, S., Tumma, A., Nikam, A., Attar, V.: SkewBoost: An algorithm for classifying imbalanced datasets. In: Proceedings of the 2nd International Conference on Computer and Communication Technology, pp. 46–52 (2011)
51. Huynh, T., Nibali, A., He, Z.: Semi-supervised learning for medical image classification using imbalanced training data. Computer Methods and Programs in Biomedicine p. 106628 (2022)
52. Johnson, J.M., Khoshgoftaar, T.M.: Survey on Deep Learning with class imbalance. Journal of Big Data **6**(1), 1–54 (2019)
53. Kaur, H., Pannu, H.S., Malhi, A.K.: A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Computing Surveys **52**(4), 1–36 (2019)
54. Khan, S.H., Hayat, M., Bennamoun, M., Sohel, F.A., Togneri, R.: Cost-sensitive learning of deep feature representations from imbalanced data. IEEE Transactions on Neural Networks and Learning Systems **29**(8), 3573–3587 (2017)
55. Khoshgoftaar, T.M., Golawala, M., Van Hulse, J.: An empirical study of learning from imbalanced data using Random Forest. In: Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence, vol. 2, pp. 310–317 (2007)
56. Kim, K.H., Sohn, S.Y.: Hybrid Neural Network with cost-sensitive Support Vector Machine for class-imbalanced multimodal data. Neural Networks **130**, 176–184 (2020)
57. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence **5**(4), 221–232 (2016)
58. Krawczyk, B., Galar, M., Jeleń, L., Herrera, F.: Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. Applied Soft Computing **38**, 714–726 (2016)
59. Krawczyk, B., Woźniak, M., Schaefer, G.: Cost-sensitive decision tree ensembles for effective imbalanced classification. Applied Soft Computing **14**, 554–562 (2014)
60. Kübler, S., Liu, C., Sayyed, Z.A.: To use or not to use: Feature selection for sentiment analysis of highly imbalanced data. Natural Language Engineering **24**(1), 3–37 (2018)
61. Kukar, M., Kononenko, I., et al.: Cost-sensitive learning with neural networks. In: ECAI, vol. 15, pp. 88–94 (1998)
62. Lee, H., Park, M., Kim, J.: Plankton classification on imbalanced large scale database via Convolutional Neural Networks with Transfer Learning. In: Proccedings of the 2016 IEEE International Conference on Image Processing, pp. 3713–3717 (2016)
63. Lee, J., Park, K.: GAN-based imbalanced data intrusion detection system. Personal and Ubiquitous Computing **25**, 121–128 (2021)
64. Lee, T., Lee, K.B., Kim, C.O.: Performance of machine learning algorithms for class-imbalanced process fault detection problems. IEEE Transactions on Semiconductor Manufacturing **29**(4), 436–445 (2016)
65. Leevy, J.L., Khoshgoftaar, T.M., Bauder, R.A., Seliya, N.: A survey on addressing high-class imbalance in big data. Journal of Big Data **5**(1), 1–30 (2018)
66. Li, W., Chen, J., Cao, J., Ma, C., Wang, J., Cui, X., Chen, P.: EID-GAN: Generative adversarial nets for extremely imbalanced data augmentation. IEEE Transactions on Industrial Informatics (2022)
67. Li, Y., Chen, J., Tan, C., Li, Y., Gu, F., Zhang, Y., Mehmood, Q.: Application of the borderline-SMOTE method in susceptibility assessments of debris flows in Pinggu District, Beijing, China. Natural Hazards **105**, 2499–2522 (2021)
68. 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)
69. Liu, H., Zhou, M., Liu, Q.: An embedded feature selection method for imbalanced data classification. IEEE/CAA Journal of Automatica Sinica **6**(3), 703–715 (2019)
70. Liu, Y., Loh, H.T., Sun, A.: Imbalanced text classification: A term weighting approach. Expert Systems with Applications **36**(1), 690–701 (2009)
71. Liu, Y., Wang, Y., Ren, X., Zhou, H., Diao, X.: A classification method based on feature selection for imbalanced data. IEEE Access **7**, 81794–81807 (2019)
72. Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M.S., Zeineddine, H.: An experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access **7**, 93010–93022 (2019)

73. Maldonado, S., Weber, R., Famili, F.: Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines. Information Sciences **286**, 228–246 (2014)
74. Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., Malossi, C.: Bagan: Data augmentation with balancing gan. arXiv preprint arXiv:1803.09655 (2018)
75. Moreo, A., Esuli, A., Sebastiani, F.: Distributional random oversampling for imbalanced text classification. In: Proceedings of the 39th ACM SIGIR International Conference on Research and Development in Information Retrieval, pp. 805–808 (2016)
76. Oak, R., Du, M., Yan, D., Takawale, H., Amit, I.: Malware detection on highly imbalanced data through sequence modeling. In: Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, pp. 37–48 (2019)
77. Obiedat, R., Qaddoura, R., Ala'M, A.Z., Al-Qaisi, L., Harfoushi, O., Alrefai, M., Faris, H.: Sentiment analysis of customers' reviews using a hybrid evolutionary SVM-based approach in an imbalanced data distribution. IEEE Access **10**, 22260–22273 (2022)
78. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: Proceedings of the 34th International Conference on Machine Learning, pp. 2642–2651 (2017)
79. Ogura, H., Amano, H., Kondo, M.: Comparison of metrics for feature selection in imbalanced text classification. Expert Systems with Applications **38**(5), 4978–4989 (2011)
80. Padurariu, C., Breaban, M.E.: Dealing with data imbalance in text classification. Procedia Computer Science **159**, 736–745 (2019)
81. Pan, T., Zhao, J., Wu, W., Yang, J.: Learning imbalanced datasets based on SMOTE and Gaussian distribution. Information Sciences **512**, 1214–1233 (2020)
82. Pouyanfar, S., Tao, Y., Mohan, A., Tian, H., Kaseb, A.S., Gauen, K., Dailey, R., Aghajanzadeh, S., Lu, Y.H., Chen, S.C., et al.: Dynamic sampling in Convolutional Neural Networks for imbalanced data classification. In: Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval, pp. 112–117 (2018)
83. Ramadhan, N.G.: Comparative analysis of adasyn-svm and smote-svm methods on the detection of type 2 diabetes mellitus. Scientific Journal Of Informatics **8**(2), 276–282
84. Ramyachitra, D., Manikandan, P.: Imbalanced dataset classification and solutions: A review. International Journal of Computing and Business Research **5**(4), 1–29 (2014)
85. Reza, M.S., Ma, J.: Imbalanced histopathological breast cancer image classification with Convolutional Neural Network. In: Proceedings of the 14th IEEE International Conference on Signal Processing, pp. 619–624 (2018)
86. Sahin, Y., Bulkan, S., Duman, E.: A cost-sensitive decision tree approach for fraud detection. Expert Systems with Applications **40**(15), 5916–5923 (2013)
87. Sampath, V., Maurtua, I., Aguilar Martin, J.J., Gutierrez, A.: A survey on Generative Adversarial Networks for imbalance problems in computer vision tasks. Journal of Big Data **8**, 1–59 (2021)
88. Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J.: Hybrid sampling for imbalanced data. In: Proceedings of the 2008 IEEE International Conference on Information Reuse and Integration, pp. 202–207 (2008)
89. Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A.: RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans **40**(1), 185–197 (2009)
90. Shelke, M.S., Deshmukh, P.R., Shandilya, V.K.: A review on imbalanced data handling using undersampling and oversampling technique. International Journal of Recent Trends in Engineering Research **3**(4), 444–449 (2017)
91. Smiti, S., Soui, M.: Bankruptcy prediction using deep learning approach based on borderline SMOTE. Information Systems Frontiers **22**, 1067–1083 (2020)
92. Soleymanpour, S., Sadr, H., Nazari Soleimandarabi, M.: CSCNN: cost-sensitive Convolutional Neural Network for encrypted traffic classification. Neural Processing Letters **53**(5), 3497–3523 (2021)
93. Song, L., Li, D., Zeng, X., Wu, Y., Guo, L., Zou, Q.: nDNA-prot: identification of DNA-binding proteins based on unbalanced classification. BMC Bioinformatics **15**, 1–10 (2014)

94. Sun, A., Lim, E.P., Liu, Y.: On strategies for imbalanced text classification using SVM: A comparative study. Decision Support Systems **48**(1), 191–201 (2009)
95. Sun, Y., Kamel, M.S., Wang, Y.: Boosting for learning multiple classes with imbalanced class distribution. In: Proceedings of the 6th International Conference on Data Mining, pp. 592–602 (2006)
96. Sun, Y., Wong, A.K., Kamel, M.S.: Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence **23**(4), 687–719 (2009)
97. Thabtah, F., Hammoud, S., Kamalov, F., Gonsalves, A.: Data imbalance in classification: Experimental evaluation. Information Sciences **513**, 429–441 (2020)
98. Ting, K.M.: An instance-weighting method to induce cost-sensitive trees. IEEE Transactions on Knowledge and Data Engineering **14**(3), 659–665 (2002)
99. Van Hulse, J., Khoshgoftaar, T.M., Napolitano, A.: Experimental perspectives on learning from imbalanced data. In: Proceedings of the 24th International Conference on Machine Learning, pp. 935–942 (2007)
100. Vo, N.H., Won, Y.: Classification of unbalanced medical data with weighted regularized least squares. In: Proceedings of the 2007 Conference on Frontiers in the Convergence of Bioscience and Information Technologies, pp. 347–352 (2007)
101. Wan, S., Duan, Y., Zou, Q.: HPSLPred: An ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source. Proteomics **17**(17-18), 1700262 (2017)
102. Wang, L., Huang, W., Lv, Q., Wang, Y., Chen, H.: AOPL: Attention enhanced oversampling and parallel deep learning model for attack detection in imbalanced network traffic. In: Proceedings of the 16th International Conference on Wireless Algorithms, Systems, and Applications, pp. 84–95 (2021)
103. Wang, Q.: A hybrid sampling SVM approach to imbalanced data classification. In: Abstract and Applied Analysis, vol. 2014 (2014)
104. Wang, S., Li, D., Zhao, L., Zhang, J.: Sample cutting method for imbalanced text sentiment classification based on BRC. Knowledge-Based Systems **37**, 451–461 (2013)
105. Wang, S., Yao, X.: Diversity analysis on imbalanced data sets by using ensemble models. In: Proceedings of the 2009 IEEE Symposium on Computational Intelligence and Data Mining, pp. 324–331 (2009)
106. Wardhani, N.W.S., Rochayani, M.Y., Iriany, A., Sulistyono, A.D., Lestantyo, P.: Cross-validation metrics for evaluating classification performance on imbalanced data. In: Proceedings of th 2019 International Conference on Computer, Control, Informatics and its applications, pp. 14–18 (2019)
107. Wei, J., Huang, H., Yao, L., Hu, Y., Fan, Q., Huang, D.: New imbalanced fault diagnosis framework based on Cluster-MWMOTE and MFO-optimized LS-SVM using limited and complex bearing data. Engineering Applications of Artificial Intelligence **96**, 103966 (2020)
108. Wei, J., Huang, H., Yao, L., Hu, Y., Fan, Q., Huang, D.: NI-MWMOTE: An improving noise-immunity majority weighted minority oversampling technique for imbalanced classification problems. Expert Systems with Applications **158**, 113504 (2020)
109. Wheelus, C., Bou-Harb, E., Zhu, X.: Tackling class imbalance in cyber security datasets. In: Proceedings of the 2018 IEEE International Conference on Information Reuse and Integration, pp. 229–232 (2018)
110. Xu, Z., Shen, D., Nie, T., Kou, Y.: A hybrid sampling algorithm combining m-smote and enn based on random forest for medical imbalanced data. Journal of Biomedical Informatics **107**, 103465 (2020)
111. Yang, H., Zhou, Y.: IDA-GAN: A novel imbalanced data augmentation gan. In: Proceedings of the 25th International Conference on Pattern Recognition, pp. 8299–8305 (2021)
112. Yang, W., Li, J., Fukumoto, F., Ye, Y.: HSCNN: a hybrid-siamese Convolutional Neural Network for extremely imbalanced multi-label text classification. In: Proceedings of the 2020 Conference on Empirical methods in Natural Language Processing, pp. 6716–6722 (2020)
113. Yang, Y., Chen, S.C.: Ensemble learning from imbalanced data set for video event detection. In: Proceedings of the 2015 IEEE International Conference on Information Reuse and Integration, pp. 82–89 (2015)

114. Yap, B.W., Abd Rani, K., Abd Rahman, H.A., Fong, S., Khairudin, Z., Abdullah, N.N.: An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In: Proceedings of the 1st International Conference on Advanced Data and Information Engineering, pp. 13–22 (2013)
115. Yen, S.J., Lee, Y.S.: Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications **36**(3), 5718–5727 (2009)
116. Yu, H., Ni, J.: An improved ensemble learning method for classifying high-dimensional and imbalanced biomedicine data. IEEE/ACM Transactions on Computational Biology and Bioinformatics **11**(4), 657–666 (2014)
117. Zhang, C., Tan, K.C., Li, H., Hong, G.S.: A cost-sensitive deep belief network for imbalanced classification. IEEE Transactions on Neural Networks and Learning Systems **30**(1), 109–122 (2018)
118. Zhang, H., Huang, L., Wu, C.Q., Li, Z.: An effective Convolutional Neural Network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset. Computer Networks **177**, 107315 (2020)
119. Zhang, H., Jiang, L., Li, C.: CS-ResNet: Cost-sensitive Residual Convolutional Neural Network for PCB cosmetic defect detection. Expert Systems with Applications **185**, 115673 (2021)
120. Zhang, H., Li, M.: RWO-Sampling: A random walk over-sampling approach to imbalanced data classification. Information Fusion **20**, 99–116 (2014)
121. Zhang, Y., Qiao, S., Lu, R., Han, N., Liu, D., Zhou, J.: How to balance the bioinformatics data: pseudo-negative sampling. BMC Bioinformatics **20**(25), 1–13 (2019)
122. Zhao, L., Shang, Z., Zhao, L., Zhang, T., Tang, Y.Y.: Software defect prediction via cost-sensitive Siamese parallel Fully-Connected Neural Networks. Neurocomputing **352**, 64–74 (2019)
123. Zhao, X.M., Li, X., Chen, L., Aihara, K.: Protein classification with imbalanced data. Proteins: Structure, Function, and Bioinformatics **70**(4), 1125–1132 (2008)
124. Zheng, Z., Cai, Y., Li, Y.: Oversampling method for imbalanced classification. Computing and Informatics **34**(5), 1017–1037 (2015)
125. Zhou, Z.H., Liu, X.Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering **18**(1), 63–77 (2005)
126. Zhu, Y., Yan, Y., Zhang, Y., Zhang, Y.: EHSO: Evolutionary Hybrid Sampling in overlapping scenarios for imbalanced learning. Neurocomputing **417**, 333–346 (2020)