

## Review

## A comprehensive review on data-level methods for imbalanced data classification

Bahareh Nikpour<sup>a,b</sup>, Farshad Rahmati<sup>a</sup>, Behzad Mirzaei<sup>a</sup>, Hossein Nezamabadi-pour<sup>a,\*</sup><sup>a</sup> Intelligent Data Processing Laboratory (IDPL), Department of Electrical Engineering, Shahid Bahonar University of Kerman, Kerman, Iran<sup>b</sup> Department of Electrical and Computer Engineering, McGill University, Montreal, Canada

## ARTICLE INFO

## Keywords:

Imbalanced data set

Classification

Data-level methods

## ABSTRACT

Classification is one of the most important tasks in machine learning and data mining. Most of the classifiers are designed for data sets with equally distributed samples among the classes. Therefore, they encounter a problem with classifying imbalanced data in which one or more classes have much fewer samples than the others. Imbalanced data sets are prevalent in the real-world, so addressing this issue is of utmost importance. There have been many methods suggested to solve this problem showing promising results, a category of which is *data-level* methods being popular for their flexibility. In this paper, our goal is to review data-level methods comprehensively and categorize them from different perspectives. Also, to simplify doing future research in this field, most of the available benchmark imbalanced data sets, software, and toolboxes are introduced. Finally, existing challenges and future works are elaborated.

## 1. Introduction

In real-world data sets, many diverse distributions of samples among different classes can be observed. In some cases, rare events result in a small number of samples in data set. For example, when the goal is to detect cancer, their number is very small compared to other cases. Another example can be fraud detection in credit transactions in which fraud cases are very less than normal ones. In the field of machine learning and data mining, such data sets are called imbalanced. In other words, imbalanced data sets are those in which data samples are unequally distributed among classes, and one or some classes have much fewer samples than the others. In binary cases, i.e., where there are two classes, the smallest class is called minority or positive class, while the largest one is the majority or negative class.

To measure the imbalance degree of data sets, there is a metric called Imbalance Ratio (IR) (Orriols-Puig & Bernadó-Mansilla, 2009) that is calculated by dividing the number of samples in the majority class,  $N_{maj}$ , to the number of minority class samples,  $N_{min}$ , as:

$$IR = \frac{N_{maj}}{N_{min}} \quad (1)$$

Usually, classification of minority class samples is a challenging task. That is while misclassifying them leads to significant risk because of their importance. On the other hand, in classifying imbalanced data, classical classifiers encounter a problem as they internally consider an even distribution for data samples. That is why imbalanced data classification has turned into an important field of research for years. Also, almost all of the data sets existing in the real-world environment are imbalanced; therefore, despite having been widely discussed in the last two decades, there are still many challenges that need to be addressed, and imbalanced data classification is still an open research field (Haixiang et al., 2017).

In recent years, many types of researches have been done to solve imbalanced data problems, and there exist several categorizations for them. A common one is to group them as Data level methods, Algorithmic level methods, and ensemble learning approaches. Data level methods aim is to preprocess the data so that the distribution of samples among classes is balanced at the end. In other words, such methods rebalance the distribution of classes using different ways of resampling the data space to reduce the inequality of the classes (Haixiang et al., 2017).

Another category includes algorithmic level approaches. In these methods, either new algorithms are developed, or the existing base

\* Corresponding author at: Intelligent Data Processing Laboratory (IDPL), Department of Electrical Engineering, Shahid Bahonar University of Kerman, Kerman, Iran.

E-mail addresses: [Bahareh.nikpour@mail.mcgill.ca](mailto:Bahareh.nikpour@mail.mcgill.ca) (B. Nikpour), [rahmatifarshad@eng.uk.ac.ir](mailto:rahmatifarshad@eng.uk.ac.ir) (F. Rahmati), [b.mirzaei@eng.uk.ac.ir](mailto:b.mirzaei@eng.uk.ac.ir) (B. Mirzaei), [Nezam@uk.ac.ir](mailto:Nezam@uk.ac.ir) (H. Nezamabadi-pour).

<https://doi.org/10.1016/j.eswa.2025.128920>

Received 5 March 2021; Received in revised form 22 November 2024; Accepted 3 July 2025

Available online 5 July 2025

0957-4174/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

classifiers are modified to acquire the ability to deal with imbalanced data sets (Batista et al., 2004; Rahmati et al., 2020). One practical approach of this category is assigning distinct costs to the training samples, so a cost matrix defines the penalties of misclassifying the samples. Such methods are called cost-sensitive, and various studies have been done regarding cost-sensitive approaches in the imbalance domain (Fernández, García, Galar, et al., 2018a; Y. Sun et al., 2007; C. Zhang, Tan, et al., 2019). Most of the algorithmic level methodologies are designed based on Support Vector Machine (SVM) (Y. Tang et al., 2009), Artificial Neural Networks (ANN) (Buda et al., 2018), Decision trees (Hoens et al., 2012), and k Nearest Neighbor (k-NN) Algorithm (Nikpour et al., 2017; Saryazdi et al., 2018). In some cases, Algorithmic level approaches take advantage of preprocessing techniques as well.

Ensemble learning is another group of techniques that emerged to improve classification performance by combining several different classifiers' decisions. These methods are widely used in the machine learning field and in different domains. Imbalanced learning is one of the fields ensemble methods have entered. As these methods are accuracy-oriented, they are combined with algorithmic and data-level methods to reach more promising results. Ensemble techniques are usually more computationally complex than the previously mentioned methods (Galar et al., 2012).

Methods of each category mentioned above have different positive and negative points, so based on the application at hand, an appropriate method should be selected (López et al., 2013). For example, algorithmic level methods are based on a specific classifier. While preprocessing techniques can be used for any data set, the resultant data can be fed to any classifier.

Another perspective to look at imbalanced learning can be a label perspective, i.e., whether the algorithms are supervised, semi-supervised, or unsupervised. In supervised learning, all training samples are labeled, while in unsupervised learning, there is no information about the samples' label. When unlabeled data cannot be discriminative and obtaining labels for all train samples is either inefficient or impossible, only a few samples will be labeled. Learning from such a data set is called semi-supervised learning. The skewed distribution causes a problem in classification (supervised learning) and clustering (unsupervised), and semi-supervised learning. Despite the importance, few works have been done for unsupervised and semi-supervised imbalanced learning, including (Krawczyk, 2016).

Several papers and books have been published to review imbalanced learning algorithms, most of which focus on reviewing methods of all categories (Abd Elrahman & Abraham, 2013; Ali et al., 2015; Bekkar & Alitouche, 2013; Branco et al., 2016; Chawla, 2009; Fernández et al., 2011; Fernández, García, Galar, et al., 2018b; Fernández, García, Herrera, et al., 2018; Haixiang et al., 2017; Krawczyk, 2016; Longadge & Dongre, 2013; López et al., 2013; Menardi & Torelli, 2014; Ramyachitra & Manikandan, 2014; Y. Sun et al., 2009; Visa & Ralescu, 2005). As a result, they do not deeply go through the methods of a single group. Also, there is a review paper that aims to study ensemble learning approaches (Galar et al., 2012). There is no paper concentrating on reviewing data-level methods to the best of our knowledge, although they are very flexible and widely used. Moreover, as mentioned before, one main advantage of such approaches is that any classifier can be used after preprocessing the data. Therefore, this paper's primary goal is to review this category's methods in more detail and provide a comprehensive taxonomy for them. In the meantime, some important issues in the imbalance learning domain are discussed as well. In summary, our goals are as follows:

- Discuss the practical applications of imbalanced data across domains like healthcare, finance, cybersecurity, and energy, categorizing them into thematic areas to emphasize the wide-ranging impact of imbalanced data challenges.
- Present a comprehensive taxonomy of data-level methods from multiple perspectives, offering a unique structured framework for

understanding and comparing approaches, including the balancing perspective, interaction with learning algorithm perspective, class perspective, inclusiveness perspective, resampling in ensembles perspective, and methods for deep learning.

- Provide a comprehensive review of the latest data-level methods, highlighting recent advancements and innovations.
- Provide a comprehensive list of publicly available datasets, software tools, and benchmarks, facilitating the replication of experiments and the exploration of future methods.
- Present a detailed analysis of evaluation metrics for imbalanced data, highlighting traditional and domain-specific measures to guide robust and fair performance assessment.
- Identify critical challenges in imbalanced data classification, such as deep learning-specific issues, and provide future research directions, offering a comprehensive framework to guide advancements in the field.

The paper is organized as follows: In section 2, the real-world applications in which data sets are imbalanced, are reviewed to reveal the importance of researching this field. Section 3 suggests a comprehensive taxonomy for data-level methods and categorizes them from different perspectives. In section 4, most of the state-of-the-art and recently proposed data level approaches are reviewed. The benchmarks and popular publicly available imbalanced datasets and available software and toolboxes designed for imbalanced learning are introduced in section 5. In section 6, the evaluation metrics are presented. In Section 7, we will discuss all existing challenges in this field and suggest future trends. Finally, a conclusion is drawn in section 8.

## 2. Imbalanced data classification applications

In recent decades, imbalanced data has become one of the most challenging aspects of data mining (Haixiang et al., 2017; Kaur et al., 2019). This issue has garnered significant attention from both academics and industry professionals due to its prevalence across various application domains (Abokadr et al., 2023; King & Zeng, 2001; Newaz & Haq, 2022; Pan et al., 2024). According to the literature, rare events occur less frequently than normal ones, yet they can have a profound impact on society (Ali et al., 2015; Ramyachitra & Manikandan, 2014).

Data analysis to predict future events based on past occurrences has gained substantial interest, playing a crucial role in automation and decision-making processes. Predicting rare events falls within this realm (Bhatta & Dang, 2023; K. He et al., 2024). These events can manifest in numerous forms, such as geohazards (e.g., tsunamis, volcanic eruptions, earthquakes, and solar flares) (Al Banna et al., 2020; Hoque et al., 2020; X. Wang et al., 2020), human hazards (e.g., fraud detection, fault monitoring, and accident prevention) (Ali et al., 2022; Pourroostaei Ardakani et al., 2023), and medical conditions (e.g., disease diagnosis) (Ahsan et al., 2022).

We have categorized imbalanced data applications into eight categories based on their thematic areas. These categories are designed to reflect the common areas in the literature, acknowledging that imbalanced data problems can occur in many other domains. Our categories are as follows:

**Biomedical & Medical:** This category includes applications in medical diagnosis, disease prediction, and healthcare management. For instance, biomedical engineering leverages engineering principles for medical applications, such as disease detection and drug resistance prediction (Ahsan et al., 2022; Grabec et al., 2019; W. Han et al., 2019; J. Lin et al., 2005; C. Sun et al., 2018).

**Security Management:** This encompasses protecting organizational assets, risk detection, and emergency management. Key tasks include risk analysis and emergency response planning (Hegde & Rokseth, 2020; Linardos et al., 2022).

**Financial Management**

Applications here include fraud detection in financial transactions and credit scoring (Abd El-Naby et al., 2023; Z. Huang et al., 2024; J. Sun et al., 2020; Throckmorton et al., 2015; J. Yao et al., 2018).

**Electrical Engineering:** This involves applications like fault detection in power systems and predictive maintenance (Ibrahim et al., 2020; Ouadah et al., 2022; Theissler et al., 2021; Y. Zhang et al., 2022).

**Information Technology:** This category includes cybersecurity and network intrusion detection systems (Farsi et al., 2018; Kocher & Kumar, 2021; Sarker et al., 2020; Shah, 2021).

**Industry:** Applications in industrial settings such as quality control and predictive maintenance (Achouch et al., 2022; Dalzochio et al., 2020; Gu et al., 2023; Niu et al., 2021).

**Energy Management:** This covers the management and prediction of energy consumption and renewable energy integration (Lai et al., 2020; Olu-Ajayi et al., 2022; Pham et al., 2020; Z. Yao et al., 2023).

**Social Services:** Applications here involve the analysis of social issues, such as public health surveillance and emergency response (Devaraj et al., 2020; Gupta & Katarya, 2020; A. A. Khan et al., 2023; Malekloo et al., 2022).

Given the primary purpose of this study is not to exhaustively detail every application of imbalanced data classification, we have focused on elaborating two widely studied categories: Biomedical & Medical and Security Management.

- **Biomedical & Medical:** Biomedical engineering (BME), also known as medical engineering, is a field of engineering tended to fill the gap between engineering and medicine. To do so, BME takes advantage of engineering principles and design concepts for medicine and biology using advanced health care like medical diagnosis (Nayak et al., 2019) and therapy (Staňková et al., 2019). This process includes designing a decision system for detecting and predicting abnormal structure for disease diagnosis (Grabec et al., 2019; W. Han et al., 2019), early warning systems for disease outbreaks (Farooq et al., 2022; Halpern, 2018), and drug resistance (Aljeldah, 2022; Nikolaou et al., 2018).
- **Security Management:** Security management is the procedure of protecting an organization's assets, including people, buildings, machines, and systems. Organizations use security management procedures to implement adequate controls on their systems. Potential risk detection, risk analysis (Azaria et al., 2014; Jomthanachai et al., 2021), and intellectual property protection are the most critical security departments' tasks. Emergency management is another topic, which stays in the security management category. The emergency is defined as the situation in which routine procedures are interrupted, and immediate measure is needed to avoid the situation from turning into a disaster that is even harder to recover from (Anderson & Adey, 2012). Emergency management is defined as managing resources for dealing with all humanitarian aspects of emergencies to reduce the harmful effects of all hazards (D. Huang et al., 2021; S. Kim et al., 2016; Maalouf & Siddiqi, 2014; R. Zhu et al., 2021).

It should be noted that this categorization aims to represent the most common applications found in the literature, recognizing that imbalanced data challenges are prevalent across numerous other domains not explicitly listed here.

### 3. Taxonomy of data level methods

A very common categorization for data-level methods in an imbalanced domain is to divide them from a Balancing perspective. As a result, we have Over-sampling, Under-sampling, and Hybrid methods. However, data-level methods can be classified from several other perspectives, including Interaction with the learning algorithm perspective,

Inclusiveness perspective, Class perspective, Resampling in ensembles perspective, and Methods for Deep Learning. Each perspective and its associated categories are explained in the following. Also, the taxonomy is summarised and illustrated in Fig. 1.

#### 3.1. Balancing perspective

As was mentioned before, data-level methods' main goal is to re-sample the data sets to become balanced at the end (Haixiang et al., 2017). There exist three main categories in this group, including Over-sampling, Under-sampling, and Hybrid methods, which are described in more detail as follows:

##### • Over-sampling

To make the classes' distribution equal, over-sampling techniques increase the number of minority class samples by duplicating the existing ones or generating new synthetic samples. The disadvantage of these methods is that they produce samples that are not real and may give the wrong perception of the data (Chawla et al., 2002). Also, the classifier's performance may be very good for training data and bad for test data after adding artificial samples, either synthetic or repeated, a problem known as overfitting (Mirzaei et al., 2021; Nikpour & Nezamabadi-pour, 2019). This problem can happen if the synthetic samples generated by an over-sampling algorithm are too similar to each other or to the existing minority class samples, thus failing to introduce sufficient variability into the training set. This lack of variability can lead the classifier to memorize the training data and perform poorly on the test data, especially in the presence of complex, and high-dimensional data (Branco et al., 2016; H. He & Garcia, 2009). Moreover, increasing the number of samples increases the complexity and the classifier's training time. Based on the previously done experiments, over-sampling techniques are more proper for data sets with a high imbalance ratio (Barandela et al., 2004; H. Yu et al., 2013).

##### • Under-sampling

Under-sampling techniques try to balance the data sets by eliminating several majority samples (Haixiang et al., 2017). This removal can be done randomly or based on different smart rules. The main drawback of methods in this category, which involve the removal of samples, is that by discarding data, potentially valuable information can be lost. This loss of information can negatively affect the classifier's ability to generalize from the training data, potentially leading to a problem called underfitting. In other words, when under-sampling is used to balance class distributions, it can indeed lead to a loss of important information if significant representatives of the classes are removed. This reduction in data can cause the model to be less complex and potentially not complex enough to capture the essential variations in the data, which can lead to underfitting. This problem is typically characterized by poor performance on both the training and test sets due to the model's inability to capture the data's underlying structure (Mirzaei et al., 2021; Nikpour & Nezamabadi-pour, 2019). As a result, it is essential to remove samples with the lowest information in the data set.

On the other hand, removing samples may reduce the training set's complexity, which can also lead to overfitting, especially if the reduced dataset is not representative of the overall data distribution. In such cases, the classifier might perform well on the training data but fail to generalize to unseen data, indicating overfitting (H. He & Garcia, 2009; Ying, 2019). Thus, while sample removal aims to address issues related to class imbalance, it must be carefully balanced to avoid the risks of both underfitting and overfitting. However, under-sampling techniques decrease the complexity and run time of classification algorithms, which is their advantage. Also, they may remove outliers and noise samples so that the performance of classifiers improves. The literature claims that

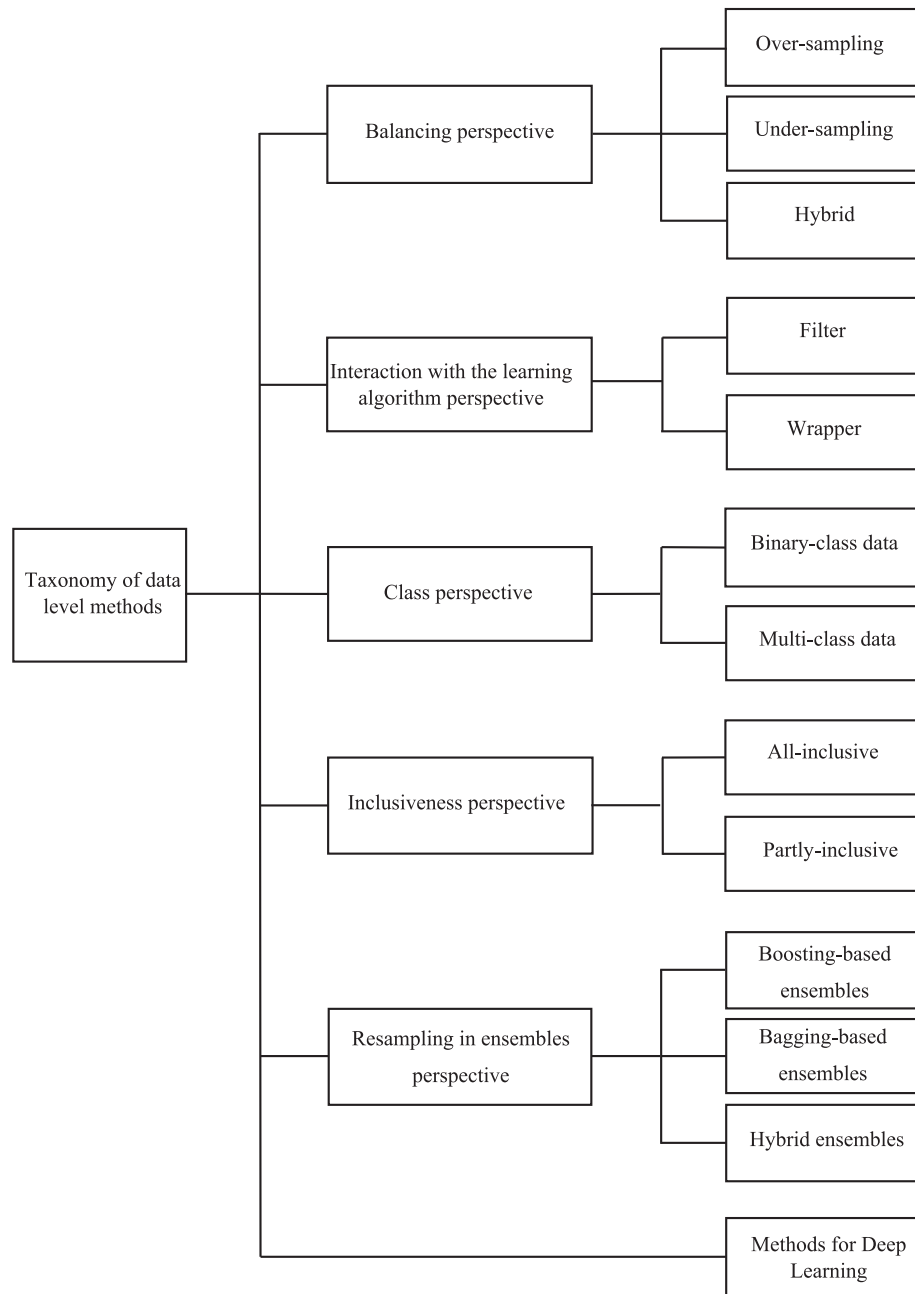


Fig. 1. Taxonomy of the data level methods.

under-sampling methods are well suited to data sets with a low imbalance ratio (Barandela et al., 2004; H. Yu et al., 2013).

Instance selection methods are considered a special case of under-sampling methods in the classification of imbalanced datasets, as they focus on reducing dataset size by removing redundant or noisy samples. These methods aim to select a representative subset of the current dataset while still fulfilling the application's original goals. The advantages of instance selection are twofold: firstly, due to the increasing amount of data generated in different fields, many large data sets exist, which leads their process to be so computationally complex; hence instance selection can be an excellent solution to reduce this complexity (Hart, 1968). Secondly, many data samples may confuse the classifier and deteriorate its performance; therefore, removing redundant samples can help to stop this occurrence and increase the interpretability and precision (García-Pedrajas, 2011). Although such methods are not explicitly designed for imbalanced data sets, they can still be useful in

this field since redundant and noisy samples may exist in both minority and majority classes. Unlike traditional under-sampling methods, which primarily aim to reduce the size of the majority class, instance selection strategically evaluates and removes redundant or noisy samples from all classes. This targeted approach helps preserve the representativeness of the dataset while enhancing its quality. Although instance selection does not pay attention to the imbalance ratio, there is a high probability that this value decreases by removing redundant samples (de Haro-García & García-Pedrajas, 2011).

#### • Hybrid

Sometimes, a combination of over-sampling and under-sampling methods may lead to better performance than applying them separately. For example, over-sampling can be applied first, and then the data set is under-sampled to reduce overfitting. There are many hybrid

methods proposed in the imbalanced domain, having shown promising results (Ramentol et al., 2012).

### 3.2. Interaction with the learning algorithm perspective

Data level approaches can also be categorized based on the method's interaction with the learning algorithm as filter and wrapper.

- **Filter**

There is no interaction with the learning algorithm in filter methods, i.e., there are some defined rules based on which samples are selected, ranked, or generated independently of the classifier. At last, the new set is used for training the classifier. Such methods are usually fast and efficient, so they are suitable for large scale and big imbalance data sets (H. Yu et al., 2013).

- **Wrapper**

In wrapper techniques, the algorithm selects or generates samples and interacts with the learning algorithm to evaluate the selected subset. It tries to make the performance better by changing the data set and receiving feedback repeatedly until the desired performance is achieved. This procedure is time-consuming, but the performance is better than the filter methods most of the time unless overfitting happens. To avoid overfitting, the train data can be partitioned to train and validation sets; then classifier is trained using train data and tested and fine-tuned on the validation set (H. Yu et al., 2013).

This perspective helps researchers decide whether to prioritize computational efficiency, as with filter methods, or to achieve higher performance through interactive optimization, as with wrapper methods.

### 3.3. Inclusiveness perspective

Based on whether the algorithm is designed to be performed on the whole data set or a section of the data, data-level methods can be categorized as follow:

- **All-inclusive**

By all-inclusive, we mean that the data level method considers the whole data samples for the preprocessing, and the locality of each sample is ignored. Most of the primary methods fall under this category.

- **Partly-inclusive**

Methods of this category first partition the data into several sections with similar characteristics and consider each section separately. The main idea is that various clusters exist in datasets, with each cluster having distinct characteristics. Cluster-based sampling algorithms fall under this category, which has received special attention from the community in recent years. First, these methods use a clustering algorithm to partition the data set and then perform under-sampling and over-sampling methods. Clustering can help represent the sampling data space better, and as a result, the local information can be seen, which helps increase the performance (Yen & Lee, 2009).

The inclusiveness perspective is particularly useful for understanding how methods differ in terms of scope. For example, all-inclusive methods may be suitable for datasets where the class distribution is uniformly imbalanced across the feature space. In contrast, partly-inclusive methods excel when the imbalance is localized within certain clusters or regions of the dataset.

### 3.4. Class perspective

From a class perspective, data-level methods are categorized into two classes of binary class and multi-class, which are explained below:

- **Methods for Binary-class data**

Most of the data level methods in the imbalanced domain are designed for binary data sets, i.e., the data sets with only two classes, but some claim that their method can be extended to multi-class data sets. Generally, binary classification algorithms can be used in multi-class data sets using two approaches: "one vs. one" and "one vs. all". In the "one vs. one" approach, each class is considered versus one of the other classes each time, and the results are aggregated at the end using a voting approach. On the other hand, in the "one vs. all" approach, each class is considered versus all other classes as a unique class (Abdi & Hashemi, 2016).

- **Methods for Multi-class data**

Although methods designed for binary class data sets can be extended to multi-class data sets, their performance may deteriorate. Therefore, a more profound vision of imbalanced learning is needed, and many issues should be considered to propose methods specified for multi-class data sets. For example, the metrics used for binary data sets cannot be applied to multi-class ones directly. A group of data-level methods focuses on addressing these issues (Abdi & Hashemi, 2016).

To find the best subset of training samples or define the best hyper-parameters of data level algorithms or classifiers, a search method should be employed to lead the procedure to explore the search space. Metaheuristic algorithms can be used for this purpose. They are swarm intelligence and evolutionary algorithms (EAs) that have been designed to find near optimum solutions for an optimization problem (Triguero et al., 2015). These methods do not promise to find global optimum, but their obtained solutions are good enough most of the time. In machine learning, metaheuristic algorithms have been used to solve feature selection, data reduction, and imbalanced data classification. In the imbalanced domain, some approaches have been proposed in data level methods that take advantage of metaheuristic and evolutionary algorithms (S. García & Herrera, 2009). As such search methods are not dependent on the number of classes, they can be used for both binary or multi-class imbalanced data sets.

One of the key challenges in multi-class datasets is the presence of long-tail distributions, where certain classes, referred to as "tail classes," contain significantly fewer samples than others, known as "head classes". This imbalance can lead to biased models that perform well on the head classes but fail to generalize on the tail classes. Traditional methods may not sufficiently address these issues in the multi-class scenarios, necessitating specialized techniques. Data-level methods, such as over-sampling "tail classes" or under-sampling "head classes", can effectively mitigate this problem (Abdi & Hashemi, 2016; Yoon & Kwek, 2005; S. Yu et al., 2022).

### 3.5. Resampling in ensembles perspective

Data level methods can be easily embedded in ensemble learning algorithms. Based on the ensemble learning algorithm used, there are three families for this group, including boosting-based ensembles, bagging-based ensembles, and hybrid ensembles, which are explained as follows (Galar et al., 2012):

- **Boosting-based ensembles**

Methods of this family embed data-level methods into boosting algorithms. In the boosting approach (Schapire, 1990), the whole data set is used to train each classifier serially. However, after each round, more



focus is given to difficult instances to correctly classify in the subsequent iteration for those classified erroneously during the current iteration. Boosting-based ensemble methods for imbalanced data change and bias the weight distribution utilized to train the next classifier toward the minority class in each iteration (Galar et al., 2012).

#### • Bagging-based ensembles

Methods of this family embed data-level methods into bagging algorithms. The bagging approach (Breiman, 1996) includes training different classifiers with bootstrapped replicas of the original training dataset. Therefore, resampling the different data subsets leads to diversity. The combination of data-level methods and bagging algorithms is usually more straightforward compared to their combination in boosting. The main factor in bagging-based ensemble methods is collecting each bootstrap replica, i.e., how class imbalance is dealt with to achieve a useful classifier in each iteration by considering the importance of the diversity (Galar et al., 2012).

#### • Hybrid ensembles

These methods conduct a double ensemble learning by combining both bagging and boosting with a data level method (Galar et al., 2012).

### 3.6. Methods for deep learning

Deep learning is a field in machine learning where artificial neural networks are used for representation learning. Recently, deep learning has gained much popularity due to its effective performance when trained with a large number of data and the progress in processors, and having access to large databases. Imbalanced data causes noticeable problems for such algorithms as it is shown that in such cases, the length of the gradient component corresponding to the majority class is much larger than the length of this component for minority class (Johnson & Khoshgoftaar, 2019). Therefore, the minority class gradient is dominated by the majority class gradient, which leads the weights to update toward decreasing majority class error. This usually causes the minority class error to increase, resulting in very slow convergence (Johnson & Khoshgoftaar, 2019). Although imbalanced data is a prevalent issue in the deep learning field, very few works have been done in this area. Most of them evaluate the effectiveness of existing imbalanced learning methods.

Some survey papers, such as (Johnson & Khoshgoftaar, 2019) and (Kaur et al., 2019), review the existing methods for solving imbalance problems in the deep learning field and present some experiments. The works focus on preprocessing imbalance data for deep learning methods that are very few despite its importance and can be a research field for interested scientists.

One of the key challenges in deep learning with imbalanced datasets is that models tend to become biased toward the majority class, which dominates the training process. Traditional approaches to tackle imbalance in deep learning, such as resampling methods (over-sampling or under-sampling), have shown mixed results. However, recent advancements in data augmentation methods have provided more sophisticated techniques for generating diverse training data, especially in the context of imbalanced learning.

Data augmentation, which involves creating new synthetic training samples by transforming or combining existing data, has emerged as an effective approach for balancing the distribution of classes without needing to collect additional data. This is particularly useful for deep learning models, where large amounts of diverse data are often necessary for optimal performance. Data augmentation has gained significant attention recently as it improves generalization and reduces overfitting in deep learning models trained on imbalanced datasets.

One widely adopted data augmentation method is Mixup (H. Zhang, 2017), which generates synthetic samples by linearly interpolating

between pairs of training samples and their corresponding labels. This method has been particularly effective in imbalanced learning because it helps the model generalize better to minority classes by creating synthetic samples that are diverse and challenging for the model to learn. Additionally, GAN-based approaches (Generative Adversarial Networks) have been applied to generate synthetic minority class samples. GANs learn to model the distribution of the minority class and can create realistic, high-quality data points that supplement the original dataset and provide an alternative to pure data augmentation methods (Ding et al., 2024; Douzas & Bacao, 2018; Mariani et al., 2018; Pan et al., 2024). Although data augmentation methods have shown great promise, their success depends on the type of data and the specific augmentation technique applied.

## 4. Literature review

One of the earliest and easiest attempts to perform under-sampling and over-sampling is to do them randomly. In other words, samples are removed randomly called Random Under-Sampling (RUS) or replicated randomly called Random Over-sampling (ROS). Since random over-sampling generates new samples by duplicating existing ones, it will increase the probability of overfitting. Moreover, random under-sampling may eliminate useful and informative samples, which is not favorable. To deal with the mentioned problems, researchers have proposed more sophisticated methods in recent years.

This section reviews most of the well-known and best performing data-level methods from the Balancing perspective, which is prevalent in other papers. These methods are reviewed based on the order of years they are presented. Afterward, the reviewed methods will be categorized based on the previous section's taxonomy in Table 1.

#### • Over-sampling methods

Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) is one of the most famous methods that generate new minority samples to balance the classes' distribution by interpolating several minority class samples gathered together in the same neighborhood. Indeed, SMOTE tends to bias the performance of the classifier and the learning towards the minority class. For this purpose, SMOTE identifies the  $k$ -nearest neighbors of each training sample. Next, depending on the required rate of over-sampling, this method randomly selects several neighbors from the neighborhood. Artificial samples are generated on the lines that connect the original sample and its neighbors. Although SMOTE performs well in most of the problems, it has several issues that need to be addressed. In the over-sampling problems, the probability of overgeneralization depends on the way the algorithm produces synthetic samples. Accordingly, as SMOTE does not consider any information for generating new samples, it may increase the probability of overgeneralization. Also, the lack of using the information for producing new samples could increase the likelihood of between class overlapping. Moreover, SMOTE algorithm is only usable for binary class problems with continuous feature space (Batista et al., 2004). It needs enough minority class samples for accurately estimating the probability distribution of the actual data. The mentioned issues motivate researchers to propose improved versions of SMOTE.

Ref. (Chawla et al., 2003) presents SMOTEBoost, which creates the synthetic minority class samples using the SMOTE algorithm. The new weights must be assigned proportionally to the new dataset's total number since the new samples are generated. The algorithm normalizes the samples' weights from the original dataset to constitute a distribution based on the new samples. Cluster-Based Over-sampling (CBO) is introduced to consider both inter- and intra-class imbalance problems using the  $k$ -means algorithm. Before over-sampling, CBO clusters both minority and majority samples. Then, random over-sampling is applied to all clusters belonging to the majority class, except for the largest one. In the end, each cluster of the majority class should have the same

**TABLE 1**  
Categorization of the reviewed methods.

Taxonomy perspective	Category	References
Balancing perspective	Over-sampling	(Chawla et al., 2002), (H. Han et al., 2005), (H. He et al., 2008), (Bunkhumpornpat et al., 2009), (S. Hu et al., 2009), (Nguyen et al., 2011), (Bunkhumpornpat et al., 2012), (Douzas et al., 2018), (Maclejewski & Stefanowski, 2011), (Cao & Wang, 2011), (Barua et al., 2014), (H. Zhang & Li, 2014), (López et al., 2014), (Abdi & Hashemi, 2016), (Thanathamathée & Lursinsap, 2013), (Sáez et al., 2016), (Nekooimehr & Lai-Yuen, 2016), (Charte, Rivera, Del Jesus, et al., 2015), (Douzas & Bacao, 2018), (Douzas & Bacao, 2017), (Mathew et al., 2018), (Ando & Huang, 2017), (Rivera, 2017), (S. Guo et al., 2019), (Piri et al., 2018), (X. Tao et al., 2019), (C. L. Liu & Hsieh, 2020), (Feng et al., 2019), (Z. Huang et al., 2020), (J. Ma et al., 2019), (S. Tang & Chen, 2008), (Jo & Japkowicz, 2004), (Barua et al., 2011), (Soltanzadeh & Hashemzadeh, 2021), (S. Wang & Yao, 2009), (Chawla et al., 2003), (H. Guo & Viktor, 2004), (Menardi & Torelli, 2014), (Castellanos et al., 2018), (Hensman & Masko, 2015), (Harliman & Uchida, 2018), (Buda et al., 2018), (Dablain et al., 2023), (Zhai et al., 2022), (Wei et al., 2022), (Y. Liu et al., 2023), (Asniar et al., 2022), (L. Tao et al., 2024), (Lu et al., 2024), (Mirzaei, 2024), (Pan et al., 2024), (Ding et al., 2024), (Escobar Díaz Guerrero et al., 2024)
	Under-sampling	(Kubat, 2000), (Laurikkala, 2001), (Yen & Lee, 2009), (H. Yu et al., 2013), (Ng et al., 2015), (Triguero et al., 2015), (D'Addabbo & Maglietta, 2015), (Kang et al., 2017), (W.-C. C. Lin et al., 2017), (Hart, 1968), (S. García & Herrera, 2009), (Tomek, 1976), (Mani & Zhang, 2003), (Yoon & Kwek, 2005), (H. J. Kim et al., 2016), (Barella et al., 2014), (Anand et al., 2010), (Seiffert et al., 2010), (Tahir et al., 2012), (Sundarkumar & Ravi, 2015), (Galar et al., 2013), (Barandela et al., 2003), (Mirzaei et al., 2020), (X. Y. Liu et al., 2009), (Hoyos-Osorio et al., 2021), (Dai et al., 2022), (H. Zhu et al., 2024), (Farshidvard et al., 2023), (Z. Sun et al., 2024), (Soltanzadeh et al., 2023), (Nikpour & Nezamabadi-pour, 2018), (Tsai et al., 2019), (Verbiest et al., 2014), (Ramentol et al., 2016), (G. Y. Y. Wong et al., 2018), (Nikpour & Nezamabadi-pour, 2019), (G. Y. Wong et al., 2013), (Lee et al., 2016), (S. Yu et al., 2022)
	Hybrid	(Seiffert et al., 2009), (Batista et al., 2004), (Stefanowski & Wilk, 2008), (Jeatrakul et al., 2010), (Tong et al., 2011), (Y. Liu et al., 2011), (Ramentol et al., 2012), (Q. Wang, 2014), (Verbiest et al., 2014), (Sáez et al., 2015), (Cateni et al., 2014), (Agrawal et al., 2015), (Charte, Rivera, del Jesus, et al., 2015), (

**TABLE 1 (continued)**

Taxonomy perspective	Category	References
Interaction with the learning algorithm perspective	Filter	(Sanguanmak & Hanskunatai, 2016), (Ramentol et al., 2016), (Mao et al., 2017), (F. Hu et al., 2019), (Susan & Kumar, 2019), (Cohen et al., 2006), (S. Wang & Yao, 2009), (Błaszczyński et al., 2010), (Mirzaei et al., 2021), (Vluymans et al., 2016), (G. Y. Y. Wong et al., 2018), (G. Y. Wong et al., 2013), (Mirzaei et al., 2022), (Arafa et al., 2022), (Dixit & Mani, 2023), (Vairetti et al., 2024) (Kubat, 2000), (Laurikkala, 2001), (Chawla et al., 2002), (H. Han et al., 2005), (Seiffert et al., 2009), (Batista et al., 2004), (H. He et al., 2008), (Bunkhumpornpat et al., 2009), (Stefanowski & Wilk, 2008), (Yen & Lee, 2009), (S. Hu et al., 2009), (Nguyen et al., 2011), (Y. Liu et al., 2011), (Bunkhumpornpat et al., 2012), (Douzas et al., 2018), (Ramentol et al., 2012), (Maclejewski & Stefanowski, 2011), (Cao & Wang, 2011), (Barua et al., 2014), (H. Zhang & Li, 2014), (Q. Wang, 2014), (Verbiest et al., 2014), (Sáez et al., 2015), (Cateni et al., 2014), (Agrawal et al., 2015), (Ng et al., 2015), (Abdi & Hashemi, 2016), (Jeatrakul et al., 2010), (Charte, Rivera, del Jesus, et al., 2015), (Sáez et al., 2016), (Kang et al., 2017), (Nekooimehr & Lai-Yuen, 2016), (Sanguanmak & Hanskunatai, 2016), (Ramentol et al., 2016), (Charte, Rivera, Del Jesus, et al., 2015), (W.-C. C. Lin et al., 2017), (Douzas & Bacao, 2018), (Douzas & Bacao, 2017), (Mathew et al., 2018), (Ando & Huang, 2017), (Rivera, 2017), (Mao et al., 2017), (S. Guo et al., 2019), (Piri et al., 2018), (X. Tao et al., 2019), (C. L. Liu & Hsieh, 2020), (F. Hu et al., 2019), (Feng et al., 2019), (Tsai et al., 2019), (Mirzaei et al., 2020), (Menardi & Torelli, 2014), (S. Tang & Chen, 2008), (Barua et al., 2011), (Mirzaei et al., 2021), (Soltanzadeh & Hashemzadeh, 2021), (Castellanos et al., 2018), (S. Wang & Yao, 2009), (Chawla et al., 2003), (Mani & Zhang, 2003), (Barandela et al., 2003), (Yoon & Kwek, 2005), (X. Y. Liu et al., 2009), (Anand et al., 2010), (Seiffert et al., 2010), (Tahir et al., 2012), (Galar et al., 2013), (Barella et al., 2014), (Sundarkumar & Ravi, 2015), (H. J. Kim et al., 2016), (Jo & Japkowicz, 2004), (Cohen et al., 2006), (Błaszczyński et al., 2010), (Hensman & Masko, 2015), (Harliman & Uchida, 2018), (Lee et al., 2016), (Mirzaei et al., 2022), (Dablain et al., 2023), (Zhai et al., 2022), (Wei et al., 2022), (Y. Liu et al., 2023), (Asniar et al., 2022), (L. Tao et al., 2024), (Lu et al., 2024), (S. Yu et al., 2022), (Hoyos-Osorio et al., 2021), (Dai et al., 2022), (H. Zhu et al., 2024), (Farshidvard et al., 2023), (Z. Sun et al., 2024), (Arafa et al., 2022), (Dixit & Mani, 2023), (Vairetti et al., 2024), (Mirzaei, 2024), (Pan et al., 2024), (Ding et al., 2024)

(continued on next page)

TABLE 1 (continued)

Taxonomy perspective	Category	References
Inclusiveness perspective	Wrapper	(Tong et al., 2011), (H. Yu et al., 2013), (Triguero et al., 2015), (S. García & Herrera, 2009), (H. Guo & Viktor, 2004), (Thanathamathée & Lursinsap, 2013), (López et al., 2014), (Vluymans et al., 2016), (Nikpour & Nezamabadi-pour, 2018), (Z. Huang et al., 2020), (J. Ma et al., 2019), (Susan & Kumar, 2019), (G. Y. Y. Wong et al., 2018), (Nikpour & Nezamabadi-pour, 2019), (G. Y. Wong et al., 2013), (Soltanzadeh et al., 2023), (Escobar Díaz Guerrero et al., 2024)
	All-inclusive	(Kubat, 2000), (Laurikkala, 2001), (Chawla et al., 2002), (Seiffert et al., 2009), (Batista et al., 2004), (H. He et al., 2008), (Jeatrakul et al., 2010), (Tong et al., 2011), (Y. Liu et al., 2011), (Ramentol et al., 2012), (Cao & Wang, 2011), (H. Yu et al., 2013), (H. Zhang & Li, 2014), (Verbiest et al., 2014), (López et al., 2014), (Sáez et al., 2015), (Cateni et al., 2014), (Abdi & Hashemi, 2016), (Charte, Rivera, del Jesus, et al., 2015), (Triguero et al., 2015), (Sáez et al., 2016), (Kang et al., 2017), (Sanguanmak & Hanskunatai, 2016), (Ramentol et al., 2016), (Charte, Rivera, Del Jesus, et al., 2015), (Douzas & Bacao, 2018), (Mathew et al., 2018), (Ando & Huang, 2017), (Rivera, 2017), (S. Guo et al., 2019), (Nikpour & Nezamabadi-pour, 2018), (X. Tao et al., 2019), (C. L. Liu & Hsieh, 2020), (Feng et al., 2019), (Z. Huang et al., 2020), (Susan & Kumar, 2019), (Mirzaei et al., 2020), (G. Y. Y. Wong et al., 2018), (S. García & Herrera, 2009), (Tahir et al., 2012), (Błaszczczyński et al., 2010), (Barandela et al., 2003), (S. Wang & Yao, 2009), (Galar et al., 2013), (H. Guo & Viktor, 2004), (Seiffert et al., 2010), (Chawla et al., 2003), (S. Tang & Chen, 2008), (Menardi & Torelli, 2014), (Mani & Zhang, 2003), (Anand et al., 2010), (Sundarkumar & Ravi, 2015), (Soltanzadeh & Hashemzadeh, 2021), (G. Y. Wong et al., 2013), (Castellanos et al., 2018), (Nikpour & Nezamabadi-pour, 2019), (Vluymans et al., 2016), (Hensman & Masko, 2015), (Harliman & Uchida, 2018), (Buda et al., 2018), (Lee et al., 2016), (Mirzaei et al., 2022), (Dablain et al., 2023), (Zhai et al., 2022), (Y. Liu et al., 2023), (S. Yu et al., 2022), (Dai et al., 2022), (H. Zhu et al., 2024), (Z. Sun et al., 2024), (Soltanzadeh et al., 2023), (Dixit & Mani, 2023), (Vairetti et al., 2024), (Pan et al., 2024), (Ding et al., 2024), (Escobar Díaz Guerrero et al., 2024)
	Partly-inclusive	(H. Han et al., 2005), (Bunkhumpornpat et al., 2009), (Stefanowski & Wilk, 2008), (S. Hu et al., 2009), (Nguyen et al., 2011), (Maciejewski & Stefanowski, 2011), (Q. Wang, 2014), (D'Addabbo & Maglietta, 2015), (Thanathamathée & Lursinsap, 2013), (Nekooimehr & Lai-Yuen, 2016), (Mao et al., 2017), (Piri et al., 2018), (F. Hu et al., 2019),

TABLE 1 (continued)

Taxonomy perspective	Category	References
Class perspective	Binary-class data	(J. Ma et al., 2019), (Tsai et al., 2019), (Jo & Japkowicz, 2004), (Barua et al., 2011), (Yoon & Kwek, 2005), (H. J. Kim et al., 2016), (Barella et al., 2014), (Cohen et al., 2006), (X. Y. Liu et al., 2009), (Mirzaei et al., 2021), (Bunkhumpornpat et al., 2012), (Barua et al., 2014), (Douzas & Bacao, 2017), (Douzas et al., 2018), (Yen & Lee, 2009), (Ng et al., 2015), (W.-C. C. Lin et al., 2017), (Agrawal et al., 2015), (Wei et al., 2022), (Asniar et al., 2022), (L. Tao et al., 2024), (Lu et al., 2024), (Hoyos-Osorio et al., 2021), (Farshidvard et al., 2023), (Arafa et al., 2022), (Mirzaei, 2024)
		(Kubat, 2000), (Laurikkala, 2001), (Chawla et al., 2002), (H. Han et al., 2005), (Seiffert et al., 2009), (Batista et al., 2004), (H. He et al., 2008), (Bunkhumpornpat et al., 2009), (Stefanowski & Wilk, 2008), (Yen & Lee, 2009), (S. Hu et al., 2009), (Jeatrakul et al., 2010), (Nguyen et al., 2011), (Tong et al., 2011), (Y. Liu et al., 2011), (Bunkhumpornpat et al., 2012), (Douzas et al., 2018), (Ramentol et al., 2012), (Maciejewski & Stefanowski, 2011), (Cao & Wang, 2011), (H. Yu et al., 2013), (Thanathamathée & Lursinsap, 2013), (Barua et al., 2014), (H. Zhang & Li, 2014), (Q. Wang, 2014), (Verbiest et al., 2014), (López et al., 2014), (Sáez et al., 2015), (Cateni et al., 2014), (Triguero et al., 2015), (D'Addabbo & Maglietta, 2015), (Kang et al., 2017), (Nekooimehr & Lai-Yuen, 2016), (Sanguanmak & Hanskunatai, 2016), (Ramentol et al., 2016), (W.-C. C. Lin et al., 2017), (Douzas & Bacao, 2018), (Douzas & Bacao, 2017), (Mathew et al., 2018), (Ando & Huang, 2017), (Rivera, 2017), (Mao et al., 2017), (S. Guo et al., 2019), (Piri et al., 2018), (Nikpour & Nezamabadi-pour, 2018), (X. Tao et al., 2019), (F. Hu et al., 2019), (Z. Huang et al., 2020), (J. Ma et al., 2019), (Tsai et al., 2019), (Susan & Kumar, 2019), (Mirzaei et al., 2020), (Mirzaei et al., 2021), (S. García & Herrera, 2009), (G. Y. Y. Wong et al., 2018), (X. Y. Liu et al., 2009), (Tahir et al., 2012), (Błaszczczyński et al., 2010), (Galar et al., 2013), (H. Guo & Viktor, 2004), (Seiffert et al., 2010), (Chawla et al., 2003), (S. Tang & Chen, 2008), (Menardi & Torelli, 2014), (Jo & Japkowicz, 2004), (Barua et al., 2011), (Tomek, 1976), (Mani & Zhang, 2003), (H. J. Kim et al., 2016), (Barella et al., 2014), (Anand et al., 2010), (Sundarkumar & Ravi, 2015), (Soltanzadeh & Hashemzadeh, 2021), (G. Y. Wong et al., 2013), (Cohen et al., 2006), (Castellanos et al., 2018), (Nikpour & Nezamabadi-pour, 2019), (Vluymans et al., 2016), (Mirzaei et al., 2022), (Zhai et al., 2022), (Wei et al., 2022), (Asniar et al., 2022), (L. Tao et al., 2024), (Lu et al., 2024), (Hoyos-Osorio et al., 2021), (Dai et al.,
		(continued on next page)



TABLE 1 (continued)

Taxonomy perspective	Category	References
	Multi-class data	2022), (H. Zhu et al., 2024), (Farshidvard et al., 2023), (Z. Sun et al., 2024), (Soltanzadeh et al., 2023), (Arafa et al., 2022), (Dixit & Mani, 2023), (Vairetti et al., 2024), (Mirzaei, 2024), (Pan et al., 2024) (Agrawal et al., 2015), (Ng et al., 2015), (Abdi & Hashemi, 2016), (Charte, Rivera, del Jesus, et al., 2015), (Sáez et al., 2016), (Charte, Rivera, Del Jesus, et al., 2015), (C. L. Liu & Hsieh, 2020), (Feng et al., 2019), (Barandela et al., 2003), (S. Wang & Yao, 2009), (Yoon & Kwek, 2005), (Hensman & Masko, 2015), (Harliman & Uchida, 2018), (Buda et al., 2018), (Lee et al., 2016), (Dablain et al., 2023), (Y. Liu et al., 2023), (S. Yu et al., 2022), (Ding et al., 2024), (Escobar Díaz Guerrero et al., 2024)
Resampling in ensembles perspective	Boosting-based ensembles Bagging-based ensembles Hybrid ensembles	(Chawla et al., 2003), (Seiffert et al., 2010), (H. Guo & Viktor, 2004), (Galar et al., 2013) (S. Wang & Yao, 2009), (Barandela et al., 2003), (Błaszczyński et al., 2010), (Tahir et al., 2012) (X. Y. Liu et al., 2009)
Methods for Deep Learning		(Ando & Huang, 2017), (Douzas & Bacao, 2018), (Hensman & Masko, 2015), (Harliman & Uchida, 2018), (Buda et al., 2018), (Lee et al., 2016), (Dablain et al., 2023), (Zhai et al., 2022), (Pan et al., 2024), (Ding et al., 2024), (Escobar Díaz Guerrero et al., 2024)
Metaheuristic-based methods		(H. Yu et al., 2013), (Triguero et al., 2015), (S. García & Herrera, 2009), (López et al., 2014), (Vluymans et al., 2016), (Nikpour & Nezamabadi-pour, 2018), (J. Ma et al., 2019), (Susan & Kumar, 2019), (Orriols-Puig & Bernadó-Mansilla, 2009), (G. Y. Y. Wong et al., 2018), (Galar et al., 2013), (H. J. Kim et al., 2016), (G. Y. Wong et al., 2013), (Soltanzadeh et al., 2023), (Nikpour & Nezamabadi-pour, 2019)

number of samples as the largest one. Eventually, all the clusters belonging to the minority class are over-sampled so that after the over-sampling process, (1) the total number of samples in the minority class equals the total number of samples in the majority class, and (2) each cluster in the minority class has the same number of samples (Jo & Japkowicz, 2004).

Authors in (H. Guo & Viktor, 2004) presented an ensemble model called DataBoost-IM that uses data generation. This algorithm identifies hard majority and minority samples during the execution of boosting. Afterward, those hard samples are selected separately and employed to generate the respective class's synthetic samples. Finally, those generated samples will be added to the main dataset. Borderline-SMOTE is one of the most well-known algorithms proposed by Han et al. (H. Han et al., 2005) to improve SMOTE. In this method, each minority class sample with a larger number of majority nearest neighbors than the number of minority nearest neighbors is defined as a borderline minority sample. Accordingly, this method considers borderline minority class samples, which are most likely to be misclassified as a set called "Danger" and tends to over-sample the examples belonging to this set. Therefore, the algorithm generates synthetic samples along the lines that connect the identified borderline samples and their nearest neighbors of

the same class. Compared to traditional SMOTE, Borderline-SMOTE enhances the robustness of the decision boundary by concentrating on regions where the minority class is most vulnerable. This targeted approach results in the generation of more relevant samples, which significantly improves classification performance.

Adaptive Synthetic Sampling (ADASYN) is another over-sampling method that tends to employ unequal distribution for different minority class samples based on their complexity level in the learning process. In this way, by decreasing class imbalance bias and shifting the decision boundary toward difficult samples, ADASYN makes the classifier more consistent with the distribution of classes (H. He et al., 2008). ADASYN is particularly effective in enhancing the detection of minority samples by focusing on those near the decision boundary. However, similar to SMOTE, ADASYN may slightly lower overall accuracy, as the increased number of synthetic samples near the boundary can lead to misclassification. In (S. Tang & Chen, 2008), an over-sampling algorithm called Adjusting the Direction Of the synthetic Minority class examples (ADOMS) is proposed. This algorithm is similar to SMOTE, but it generates synthetic samples along the first principal component axis (PCA) of the local data distribution using  $k$  nearest neighbors.

Safe-Level Synthetic Minority Over-sampling Technique (Safe-Level-SMOTE) is another method that assigns a safe level ratio to each minority class sample and generates synthetic samples only in the safe regions. Contrary to SMOTE and Borderline-SMOTE, which may generate samples in inappropriate regions, Safe-Level-SMOTE's idea prevents the generation of samples that can increase overlapping and noise (Bunkhumpornpat et al., 2009). It should be noted that Safe-Level-SMOTE serves as a complementary method to ADASYN. While ADASYN addresses the risk of overfitting by focusing on difficult minority instances, Safe-Level-SMOTE is designed to prevent underfitting by ensuring that synthetic samples are generated only in 'safe' regions of the feature space, where the likelihood of class overlap is minimized.

MSMOTE is another modification on SMOTE in which minority class samples are categorized into three groups: security samples, border samples, and noise samples based on the distance of all samples. MSMOTE takes minority class samples into account and removes noise samples by adaptive mediation (S. Hu et al., 2009). The OverBagging procedure proposed by Wang et al. (S. Wang & Yao, 2009) increases the size of the minority class by iterating over the original samples (random over-sampling), whereas the majority class samples can be all considered in every bag or can be resampled to increase the diversity. That is while SMOTEBagging integrates bagging with different amounts of SMOTE and over-sampling in each iteration so that the dataset is balanced completely. Over-sampling percentage in this method changes in each iteration (ranging from 10 % in the first iteration to 100 % in the last, and is multiple of 10). This ratio shows the number of minority samples resampled randomly (with replacement) from the original dataset in every iteration. The rest of the minority samples will be created by SMOTE (S. Wang & Yao, 2009).

Ref. (Nguyen et al., 2011) presents a new over-sampling method by combining extrapolation and interpolation techniques. Extrapolation expands the minority class boundaries toward areas with fewer majority class instances. Meanwhile, the current rare class boundary is consolidated by interpolation to other places. One problem of original SMOTE is that it overgeneralizes the minority class area as it does not take the distribution of neighbors from the majority classes into account. LN-SMOTE, presented in (Maciejewski & Stefanowski, 2011), tries to solve this problem by adapting and modifying the idea of identifying the distribution of other examples located in the neighborhood. Experimental results reveal that LN-SMOTE outperforms the original SMOTE and borderline version of it. ASMOBD (over-sampling technique based on data density) is an over-sampling technique that takes advantage of the data's density to generate new minority samples (Cao & Wang, 2011). Compared to other existing methods, ASMOBD can generate a different number of synthetic samples around each rare class sample based on its difficulty level in the learning process. Also, this method can

be used to eliminate noise from the dataset. It should be noted that ADASYN can be considered a specialized variant of ASMOBD, as both methods prioritize generating synthetic samples based on the difficulty level of minority class instances.

Authors in (Barua et al., 2011) proposed a Cluster-Based Synthetic Over-sampling (CBSO) algorithm, which combines the synthetic over-sampling mechanism of existing methodologies with a different data generation mechanism based on clustering. CBSO creates the synthetic data samples using an unsupervised clustering method rather than the  $k$ -NN strategy and ensures that these samples always lie inside the minority regions. DBSMOTE is a cluster-based over-sampling technique that tries to form arbitrarily shaped clusters using the DBSCAN method (Ester et al., 1996). In the next step, DBSMOTE generates artificial samples along the shortest path between each minority sample and a pseudo centroid of a minority-class cluster. As a result, the generated samples are dense near centroids and sparse getting far from them (Bunkhumpornpat et al., 2012). In (Thanathamathee & Lursinsap, 2013), the authors proposed a method for handling imbalanced data problems based on two concepts. The first concept is to show the effect of measuring the distance between classes' subclusters and demonstrate all relevant class boundary samples. The second one is expanding the distribution of training space to deal with unseen incoming testing instances in advance based on Bootstrapping methods.

Majority Weighted Minority Over-sampling Technique (MWMOTE), introduced by Barua et al., takes advantage of two steps for producing synthetic samples. First, it identifies hard-to-learn informative minority class samples and assigns them weights based on their Euclidean distance from the nearest majority class samples. Next, artificial samples are generated based on the weighted informative minority class samples using a clustering approach. Hence, all the generated synthetic samples lie inside the minority class clusters (Barua et al., 2014). A random walk over-sampling method, RWO-sampling, is presented in (H. Zhang & Li, 2014), which uses random walking to generate new minority samples. The distribution of new synthetic samples follows the primary minority samples, and the boundary of this class is expanded. Iterative Instance Adjustment algorithm for Imbalanced Domains is another algorithm that learns the proper number of samples to represent the classes iteratively (López et al., 2014). This method employs an evolutionary algorithm to optimize the samples' placing and select the best samples to demonstrate the classes. Random Over-sampling Examples (ROSE) is a smoothed bootstrap-based technique to mitigate the effect of the extreme imbalanced distribution of classes. ROSE creates the synthetic samples from the conditional density estimate of the classes. This method benefits model estimation and assessments and applies to both continuous and categorical datasets (Menardi & Torelli, 2014).

There are data sets associated with more than one label in the real world, called multi-label data sets, some of which are imbalanced (Kashef et al., 2018). In (Charte, Rivera, del Jesus, & Herrera, 2015; Charte, Rivera, Del Jesus, & Herrera, 2015; Fernández, García, Herrera, & Chawla, 2018; Kim, Jo, & Shin, 2016; Prati, Batista, & Monard, 2004; Zhang, Bi, et al., 2019; Zhang, Tan, Li, & Hong, 2019; Fernández et al., 2018a; Fernández et al., 2018b), SMOTE is modified to acquire the ability to be used for such data sets. This method is called MLSMOTE. In (Hensman & Masko, 2015), ROS is used to make the distribution of data balanced, and afterward, its effect on Convolutional Neural Network (CNN) was observed. The authors made an imbalanced version of the popular CIFAR dataset and found out that ROS could lead to results as good as the original balanced data results. Abdi et al. proposed an over-sampling technique based on Mahalanobis Distance called MDO for multi-class imbalanced sets in which synthetic minority samples are generated with the Mahalanobis distance from the class mean similar to the other minority samples (Abdi & Hashemi, 2016). MDO decreases the risk of between-class overlapping, which is very important in multi-class problems. An over-sampling approach for multi-class imbalanced data sets is proposed in (Sáez et al., 2016). This approach focuses on using the characteristics of classes so that in each class, subsets of specific samples

are found, and over-sampling is arranged independently for each of them to use class structure information. A novel over-sampling technique is presented in (Nekooimehr & Lai-Yuen, 2016) that defines borderline minority samples by clustering the minority class. The clustering methodology is a hierarchical semi-unsupervised algorithm. Then, it defines the over-sampling size for each cluster adaptively and over-samples minority samples based on their distance to majority samples. In this way, majority and minority classes will not overlap since majority samples are considered in clustering and over-sampling processes.

The self-Organizing Map-based Over-sampling method (SOMO) employs a self-organizing map to present the input space in two dimensions, making the production of new samples more effective. After mapping the input space, SOMO creates within-cluster and between-cluster synthetic samples (Douzas & Bacao, 2017). Deep Over-sampling (DOS) was proposed by Ando and Huang, where the synthetic over-sampling method was extended to deep feature space gained by the convolutional neural network (CNN) (Ando & Huang, 2017). They used an iterative process to train CNN, which results in reducing the in-class variance. Noise Reduction A Priori Synthetic Over-Sampling methodology is proposed in (Rivera, 2017), which considers a membership probability for minority samples. Therefore, it can eliminate minority samples that appear to be noise, and the sampling will be selective. Most of the techniques that are presented in the SMOTE category are complex and increases noise during over-sampling. Accordingly, Douzas et al. propose a simple and effective over-sampling method based on  $k$ -means clustering and SMOTE, which avoids generating noise and effectively deals with inequality between and within classes' distributions (Douzas et al., 2018). Unlike most data-level methods, which are designed for data being represented by a feature vector, a method was proposed in 2018 to balance data sets in string space (Castellanos et al., 2018). In this method, SMOTE is adapted to string space, i.e., synthetic strings are generated between two training samples.

To solve the problem SMOTE encounters in nonlinear problems, weighted kernel-based SMOTE is proposed that generates new samples in the support vector machine's feature space (Mathew et al., 2018). In (Douzas & Bacao, 2018), the conditional version of Generative Adversarial Networks is used to estimate the actual distribution of data and generate artificial minority samples, so it is considered an over-sampling technique and designed for binary data sets. The synthetic informative minority over-sampling method is introduced in (Piri et al., 2018), which takes advantage of SVM's decision boundary. This method first classifies the data using SVM and finds the decision boundary. Afterward, the samples close to the decision boundary are over-sampled. Another version of this method is also designed that gives more weight to incorrectly classified samples and over-samples them with a higher degree.

An approach for solving the imbalanced data problem in deep learning was presented in (Harliman & Uchida, 2018) which uses both over-sampling and algorithm modification technique. In this method, a modified version of SMOTE is proposed for over-sampling the minority class. They called their method Ripple-SMOTE, which generates synthetic samples using  $i$ -farthest samples from centroids. The authors used MNIST, CUREt texture set, and Malware data set for their experiments and proved that their approach could improve performance. In another study in deep learning domain (Buda et al., 2018) proved that ROS outperforms RUS and two-phase learning methods by doing experiments on three large data sets, including MNIST, CIFAR-10, and ImageNet, and comparing the AUC results.

In (S. Guo et al., 2019), an improvement for SMOTE is proposed. The Euclidean distance of each minority sample is employed to regulate classes' distribution and create artificial samples in the other minority samples' neighborhoods. Xinmin et al. introduced a new over-sampling method in 2019 that creates synthetic data using a real-valued negative selection process without needing original minority samples (X. Tao et al., 2019). The synthetic samples and rare original minority samples

are then combined with majority samples to generate a balanced data set. A dynamic rotation forest algorithm was designed for multi-class imbalanced data sets based on the SMOTE algorithm (Feng et al., 2019). The idea was to balance the data sets before constructing rotation decision trees. A number of desired minority samples are replicated based on a defined rate, and the rest are generated using the SMOTE algorithm.

Huang et al. tried to improve the MOTE algorithm by making it dynamic in (Z. Huang et al., 2020). They proposed an algorithm to optimize the SMOTE parameter and the final imbalanced ratio selected by the user before. To do so, they recommended a new selection strategy and used a state transition algorithm to search for the best solutions. An evolutionary over-sampling technique based on safe-level approach is proposed in (J. Ma et al., 2019) that generates synthetic samples along the same line but with different weights. This method is used to predict seminal quality. In (C. L. Liu & Hsieh, 2020), a framework is suggested that combines sampling and modeling methods for data generation. This method, called model-based synthetic sampling, utilizes regression to find the connection between features and considers data generation diversity. In (Soltanzadeh & Hashemzadeh, 2021), an improved minority over-sampling method called the Range-Controlled SMOTE (RCSMOTE) is introduced to simultaneously overcome all SMOTE method problems. This method's purpose is twofold: 1) improving the sample selection procedure in the over-sampling process, and 2) improving the synthetic sample generation manner.

Zhai et al. presented two innovative methods, BDC1 and BDC2, aimed at enhancing binary imbalanced data classification through the application of diversity over-sampling techniques facilitated by generative models (Zhai et al., 2022). These approaches address the limitations of traditional over-sampling methods like SMOTE, which frequently fail to capture the comprehensive probability distribution of data and generate samples with limited diversity. BDC1 utilizes an Extreme Learning Machine Autoencoder (ELMAE) to iteratively refine the generation of minority class samples, increasing their diversity while maintaining their distinct characteristics separate from the majority class. On the other hand, BDC2 employs a Generative Adversarial Network (GAN) that fosters a dynamic adversarial process to produce high-quality synthetic samples, significantly enhancing the separability and diversity compared to conventional methods. These advanced strategies outperform existing methods by producing samples that are not only more diverse but also better positioned to delineate the decision boundary between classes, thereby improving the classifier's performance on imbalanced datasets.

Wei et al. proposed an enhanced version of SMOTE for addressing data imbalance issues, named Improved and Random SMOTE (IR-SMOTE). This technique enhances the traditional SMOTE approach by employing  $k$ -means clustering to remove noise and using kernel density estimation to adaptively generate synthetic samples, ensuring representative diversity. Designed to improve classification performance by more effectively generating synthetic samples for minority classes in imbalanced datasets, IR-SMOTE differentiates itself from traditional SMOTE by introducing randomness in the selection process of minority class examples and the interpolation method to create new instances. This approach aims to diversify the synthetic samples and avoid overfitting to specific patterns present in the minority class. By incorporating a random selection of features and examples for generating new data points, IR-SMOTE ensures a broader coverage of the feature space, which helps in capturing the underlying distribution more accurately (Wei et al., 2022). Asniar et al. introduced SMOTE-LOF method, which enhances SMOTE by integrating the Local Outlier Factor (LOF) to identify and eliminate noisy synthetic samples in imbalanced datasets (Asniar et al., 2022). While SMOTE helps to balance class distribution by generating synthetic minority samples, it can inadvertently create samples that mimic the majority class, introducing noise that can degrade model performance. SMOTE-LOF addresses this by applying LOFA technique traditionally used for outlier detection to refine the

dataset, ensuring the synthetic samples accurately represent the minority class and improve the overall predictive accuracy of the model.

DeepSMOTE is another approach introducing a novel end to end over-sampling algorithm specifically designed for deep learning models applied to image data with class imbalances (Dablain et al., 2023). The algorithm integrates the principles of SMOTE within a deep learning architecture comprising an encoder/decoder framework, a dedicated loss function enhanced with a penalty term, and SMOTE-based over-sampling. This structure allows for effective, high-quality artificial image generation, addressing the bias toward majority classes. Unlike methods that rely on GANs, DeepSMOTE does not require a discriminator component, simplifying the training process and enhancing the overall stability and robustness of the model. A noise-robust over-sampling method is discussed in (Y. Liu et al., 2023) which primarily addresses issues related to noisy and overlapping data in imbalanced datasets. The approach begins by identifying clusters within the data, where each cluster is analyzed to determine if it contains noise data points that may overlap significantly with majorities, causing confusion for classifiers. Once identified, these noisy instances are removed. Subsequently, safe and noise-free synthetic samples are generated through a method that carefully considers the proximity and density of the minority class samples, ensuring that the newly created points reinforce the minority class without exacerbating overlap or introducing new noise. This process results in a cleaner, more balanced dataset that enhances the performance of classification models.

The Adaptive Safe-Region Diversity Over-sampling (ASRDO) method is proposed in (L. Tao et al., 2024) which introduces an approach to address imbalanced classification by defining safe hyperspherical sampling regions around each minority class instance, using the nearest distance to majority instances as a radius. This prevents the encroachment of synthesized instances into majority class territory, thereby reducing noise. The method assigns weights based on the density within these regions and the distance to the nearest majority instances, prioritizing harder-to-learn minority instances for over-sampling. New synthetic instances are then generated along a direction vector, determined by linear combination of vectors from the nearest minority neighbors, ensuring that these instances are both diverse and representative of the minority class.

Lu et al. proposed the Overlapping Minimization-based Over-Sampling (OMOS) algorithm to enhance binary imbalanced classification by strategically generating synthetic minority samples to augment the dataset while minimizing overlap with the majority class (Lu et al., 2024). This process involves four key steps: first, employing the mean shift algorithm to dynamically cluster the dataset based on data density, thereby identifying natural groupings without predetermined cluster counts. Second, filtering to select 'safe' samples, those that consistently retain their minority status post-clustering to ensure the synthetic samples introduced do not incorporate characteristics of the majority class. Third, training autoencoders for each minority cluster to capture the essential distribution features of these safe samples, facilitating the generation of representative and diverse synthetic instances. Finally, adaptive over-sampling utilizes the modeled distributions to create new samples, with sampling rates calculated for each cluster based on its sparsity and sample density, ensuring a balanced enhancement of the dataset that maintains the original data integrity and improves classifier performance.

The work in (Escobar Díaz Guerrero et al., 2024) presents an over-sampling methodology designed to address class imbalance in histopathology images, specifically for nuclei detection. It introduces a modified copy-paste data augmentation technique combined with weight-balancing in the loss function to improve classification performance of minority classes without compromising majority classes. It is worth noting that while the primary focus of this work is on data-level methods, it also incorporates algorithmic adjustments to further improve performance. The approach is validated on an unbalanced dataset, demonstrating its effectiveness in enhancing detection and



classification, particularly in cases with high instance density. Ding et al. proposed a GAN-based over-sampling method, IBGAN, to address the issue of imbalanced medical image classification using data augmentation. It focuses on generating intra-class sparse and boundary samples using techniques like iForest and SVDD to enhance the quality and diversity of synthetic data, ultimately improving classification performance (Ding et al., 2024).

Ref. (Pan et al., 2024) introduces an Improved Generative Adversarial Network (I-GAN) for over-sampling imbalanced datasets, addressing the limitations of traditional methods that rely on local density distributions. I-GAN incorporates three strategies: sampling random vectors for the generator from a rough estimate of the distribution of minority samples, enhancing the discriminator's loss function, and reshaping generated samples to better reflect the minority class distribution. Experimental results show that I-GAN significantly outperforms 22 classical sampling methods and other GAN approaches, demonstrating its effectiveness in improving classification performance. The work in (Mirzaei, 2024) proposes a novel over-sampling technique based on clustering. Initially, the k-means clustering algorithm is applied to group the minority class samples. Next, the sparse clusters that contain fewer samples are selected. Finally, the synthetic samples for the minority class are generated using the nearest neighbors of each cluster center. Additionally, the roulette wheel selection operator is employed to probabilistically choose clusters during the over-sampling process.

#### • Under-sampling methods

In this category, Condensed Nearest Neighbor Rule (CNN) is used to generate a subset  $E'$  from  $E$ . Firstly, the algorithm draws one majority class sample and all the minority class samples and puts these samples in  $E'$ . Then, it applies 1-NN over the samples in  $E'$  to classify the samples in  $E$  in such a way that every misclassified sample from  $E$  is moved to  $E'$ . It should be noted that this algorithm does not find the smallest consistent subset from  $E$ , and its main aim is to remove the majority class samples that are distant from the decision border because they may be considered less relevant for learning (Hart, 1968). Tomek Links (TL) is a method that can be used as an under-sampling or as a data cleaning technique (Tomek, 1976). Let  $E_i = (x_i, y_i)$  and  $E_j = (x_j, y_j)$  denote two samples where  $y_i \neq y_j$  and  $d(E_i, E_j)$  shows the distance between  $E_i$  and  $E_j$ . A pair  $(E_i, E_j)$  is called Tomek Link if there is not a sample  $E_l$ , so that  $d(E_i, E_l) < d(E_i, E_j)$  or  $d(E_j, E_l) < d(E_j, E_i)$ . If two samples form a Tomek link, then either one of these samples is noise or both samples are borderline. As an under-sampling method, only majority class samples are eliminated, and as a data cleaning method, samples from all classes are removed in each Tomek Link found.

Kubat et al. proposed an under-sampling method called one-sided selection (OSS) in (Kubat, 2000). This method first creates a set containing all minority class samples and one randomly selected majority class sample and calls it  $C$ . Then, the original training set is classified by one nearest neighbor using set  $C$ . All misclassified samples are transferred to  $C$ . Finally, the majority class samples participating in Tomek Links are removed from  $C$ , which results in the removal of borderline and noisy samples. Neighborhood Cleaning Rule (NCL) is an under-sampling method in which, first, the three nearest neighbors of each training sample are found. If this sample is of majority class and is misclassified by its three nearest neighbors, the algorithm removes it. If the sample is a minority class sample and its three nearest neighbors classify it wrongly, the nearest neighbors belonging to the majority class are eliminated (Laurikkala, 2001).

NearMiss approaches are a family of four under-sampling methods based on informed heuristics (Mani & Zhang, 2003). The first method is "NearMiss-1", in which the majority class samples are selected close to some of the minority class samples. This methodology selects the majority samples with the smallest average distances to their three closest minority samples. The second method is "NearMiss-2", in which those

majority samples whose average distances to the three most distant minority samples are the smallest will be chosen. The third method is "NearMiss-3", which chooses a given number of the closest majority samples for each minority sample. Finally, the fourth method is called "Most distant", which selects those majority samples whose average distances to the three nearest minority samples are the largest. Authors in (Barandela et al., 2003) proposed the UnderBagging procedure, which on the contrary to OverBagging, works based on under-sampling the dataset in each bagging iteration randomly such that all the minority class samples are kept in each iteration. In (Yen & Lee, 2009), the authors propose a cluster-based under-sampling technique, called SBC, by taking advantage of a backpropagation neural network. SBC achieves not only satisfying results in identifying the rare class events but also has fast execution time.

Class Purity Maximization (CPM) is an under-sampling technique that first finds a pair of samples as centers, one belonging to the minority class and the other belongs to the majority class. These centers are used to partition all the samples into two clusters  $C_1, C_2$ . When the class impurity in either of the clusters is less than its parent impurity ( $Imp$ ), our clusters have been found. The impurity of a set of samples is simply the proportion of minority class samples. Then, each of these clusters is partitioned into subclusters recursively. Accordingly, hierarchical clustering is formed. If the impurity cannot be improved, then the recursion will be stopped (Yoon & Kwek, 2005). Evolutionary Under-sampling with the CHC algorithm (EUSCHC) is an under-sampling technique in which selecting the subset of samples is done by the well-known CHC evolutionary algorithm and considering a binary codification for the subset membership. In this method, any performance measure can be used as a fitness, and the samples classified correctly and outside of the chosen subset have positive weighting (S. García & Herrera, 2009).

Liu et al. (X. Y. Liu et al., 2009) proposed two hybrid ensemble methods named EasyEnsemble and BalanceCascade, referred to as exploratory under-sampling methodologies. EasyEnsemble is performed similarly to UnderBagging, but despite the training of a classifier for each new bag, each bag is trained using AdaBoost. Accordingly, the final classifier looks like an ensemble of ensembles, whereas it is a single ensemble. Besides, BalanceCascade is a supervised algorithm in which the classifiers have to be trained sequentially. In each bagging iteration, after learning AdaBoost, those majority samples which are classified correctly with higher certainties by the currently trained classifiers will be eliminated from the dataset, and they do not exist in further iterations. Ref. (Anand et al., 2010) presents a weighted under-sampling based approach in which first, the weighted Euclidean distance of each majority sample from each of the minority samples is calculated. All features are weighted by its Fishers discriminant score, which measures the overlapping per attribute. Afterward, for each minority sample, the majority samples are sorted in ascending order of distance from the minority sample. Finally, for each minority sample, a user-defined number of majority samples are chosen. The user-defined number denotes the desired ratio of majority samples to minority samples. Seiffert et al. (Seiffert et al., 2010) introduced RUSBoost, which eliminates the majority class samples in each iteration using random under-sampling. To form a distribution, the weights of samples in the new under-sampled dataset will be normalized.

In (Tahir et al., 2012), the authors introduced an inverse random under-sampling technique (IRUS) for class imbalance problems based on the bagging concept. The main idea of IRUS is severely under-sampling the majority class through creating bags such that the imbalanced situation is reversed concerning the original one. Each bag includes all minority samples but only a few majorities. Thus, the classification model concentrates on the minority class that can be distinguished from the majority class successfully. A combination of SMOTE and CHC evolutionary algorithm is proposed in (G. Y. Wong et al., 2013) to improve SMOTE results. This method first over-samples the minority class, and then the CHC evolutionary algorithm is used to decrease the synthetic samples and the majority samples. ACOSampling (H. Yu et al.,

2013) takes advantage of ant colony optimization (ACO) to eliminate redundant samples from the dataset. First, it starts with feature selection to remove noisy genes from the data. Next, it divides the original training set into two sets, including the training and validation set, randomly and repeatedly. Then, in each subset, a modified version of the ACO algorithm is conducted to eliminate less critical majority class samples and search for the optimal training subset. In the end, a statistical result of each training subset is given in the form of a frequency list where each frequency denotes the importance of each majority class sample in the learning procedure. Finally, highly significant majority samples are kept, and the final training set is constructed by combining them with minority samples.

Galar et al. (Galar et al., 2013) proposed an intelligent under-sampling-based ensemble named EusBoost that combines the EUS data level method in each iteration of AdaBoost. The EusBoost method works based on RusBoost and increases classifier performance by applying the evolutionary under-sampling technique. The critical factor of EusBoost is the diversity mechanism in which a different subset of data in each iteration is considered. Verbiest et al. proposed a hybrid method for noisy imbalanced data sets that first applies a prototype selection algorithm on data sets to remove noninformative or noisy samples. Then, the SMOTE algorithm is used for over-sampling, and finally, another prototype selection algorithm is used to clean the generated data. Both prototype selection algorithms are designed based on the theory of fuzzy rough sets (Verbiest et al., 2014).

Barella et al. (Barella et al., 2014) introduced an under-sampling strategy named ClusterOSS, an adaptation of the strategy used by OSS to balance the data distribution. Firstly, the algorithm uses a clustering procedure (for example,  $k$ -means) to cluster the majority class samples. Then, for each cluster, the closest samples to the center are found. These samples are employed to start the under-sampling process, which is identical to OSS. Eventually, as in OSS, Tomek Links data cleaning technique is utilized. There are two main differences between ClusterOSS and OSS. The first difference is that ClusterOSS can start the under-sampling process from more than one sample and the second one is that the under-sampling process is not started at random. By doing so, ClusterOSS can enhance the effectiveness of under-sampling. In (Ng et al., 2015), a technique is proposed which clusters the majority samples to consider the information regarding their distribution and improve the diversity. Then, the under-sampling is performed by selecting samples using the stochastic sensitivity measure. This procedure is continued iteratively until a balanced data set is achieved. A parallel model for evolutionary under-sampling in large scale imbalanced data sets is proposed in (Triguero et al., 2015). To do so, the MapReduce framework is adopted, and a windowing approach is developed to accelerate the under-sampling procedure.

Sundarkumar et al. (Sundarkumar & Ravi, 2015) present a hybrid under-sampling method in which, first, the atypical values from the majority class are eliminated using a  $k$  reverse NNs technique. Then, from the resulting dataset without outlier data, the support vectors are extracted using one-class SVMs. In this method, SVM is selected mainly, as it performs an IS (through the collection of support vectors) while classifying the data sets. Accordingly, the under-sampling process is carried out implicitly by choosing some important samples from the majority class. Parallel Selective Sampling is a filter under-sampling method designed for big imbalanced data sets (D'Addabbo & Maglietta, 2015). This method's main idea is that the samples existing close to the separating boundary are more relevant, so such samples should be well-preserved when decreasing the data size. Another SMOTE-based method called SMOTE-FRST-2 T is proposed in (Ramentol et al., 2016). This method combines SMOTE and an instance selection strategy designed based on a fuzzy rough set theory. After applying SMOTE to the data set, two thresholds are used to remove redundant majority samples and useless generated minority samples. In another study, a two-phase learning method was employed. RUS was used to under-sample the majority class, and a Convolutional Neural Network

was pre-trained using the new data (Lee et al., 2016). Afterward, CNN was fine-tuned using the original data. Their experiments on WHOI-Plankton data, a big, highly imbalanced data, proved their method is effective.

In (H. J. Kim et al., 2016), the authors propose a cluster-based evolutionary under-sampling technique, called CBEUS, in which clustering and genetic algorithms are combined to address the problem of imbalanced data. CBEUS first divides the majority class samples into several clusters using the  $k$ -means algorithm. Afterward, the distance between a sample and centroid within each cluster is calculated using the Euclidean distance function. In the next step, the thresholds representing the distance from each group's centroid are found using a genetic algorithm. In this way, the relevant majority samples are selected based on removing noisy samples that are far from the centroid of the cluster. Noise-filtered Under-sampling Scheme (NUS) is suggested in (Kang et al., 2017), which applies  $k$  Nearest Neighbor based noise filter to the minority class since noisy minority samples can deteriorate the classifier's performance. Afterward, the majority class is under-sampled using one of the four under-sampling methods analyzed in the paper called UA, RUS, UB, and EE. In (W.-C. C. Lin et al., 2017), two under-sampling approaches are presented. In both approaches, the majority class samples are clustered using the  $k$ -means algorithm, and the number of classes is equal to minority samples. In the first strategy, cluster centers represent majority class samples, while in the second strategy, the nearest neighbor to cluster centers is used.

Hyper-heuristic training set selection is another methodology that uses a metaheuristic approach called BQIGSA in a hyper-heuristic framework to select informative samples from the majority and minority classes (Nikpour & Nezamabadi-pour, 2018). This framework allows a trade-off between exploration and exploitation in the search algorithm using global and local search approaches. In (G. Y. Y. Wong et al., 2018), the authors introduced a hybrid preprocessing method that uses Fuzzy Rule Base (FRB) with CHC evolutionary algorithm called FRB + CHC. First, fuzzy logic is utilized to over-sample the minority class samples, and a fuzzy rule base is formed. The fuzzy rules are created based on the minority class samples. To evaluate the importance of each rule, the rule weight  $w_0$  is employed. After generating the rule base for the minority class, the rules are drawn randomly regarding the rule weight. The rule with a higher rule weight has a higher chance to be chosen, and a new synthetic sample is generated inside the area of the chosen rule. This process is repeated until the minority class size is the same as that of the majority class. Finally, the CHC algorithm removes both of the new synthetic samples and the majority samples.

A cluster-based instance selection algorithm is proposed, which first clusters majority samples and then selects the best samples in each subclass using instance selection (Tsai et al., 2019). A memetic algorithm for training set selection is presented in (Nikpour & Nezamabadi-pour, 2019), which selects useful samples from minority and majority classes. This method first proposes a memetic based search framework using a binary quantum-inspired gravitational search algorithm. It then suggests several local search algorithms used in this framework based on the problem's nature. In Ref. (Mirzaei et al., 2020), an under-sampling technique is presented, using the DBSCAN algorithm to reduce the majority class size. First, this technique applies the DBSCAN algorithm over the majority samples to cluster them into different categories, including core samples, border samples, and outlier samples. Then, only core samples are employed as representative of the majority class and other samples including border and outlier samples are eliminated to obtain the balanced training set.

An under-sampling method called Relevant Information-based Under-Sampling (RIUS) is presented in (Hoyos-Osorio et al., 2021) which aims to improve classification performance by selecting the most relevant examples from the majority class, based on an information-preservation principle. The method incorporates clustering-based under-sampling (CBUS) to enhance data representation, forming a more robust approach known as Clustering-based RIUS (CRIUS). RIUS



identifies informative instances using an entropy-based cost function that measures redundancy and distortion within the majority class data. CRIUS builds on this by dividing the majority class into clusters and applying RIUS within each to further optimize the sample selection, balancing the dataset without losing significant data structures. In another paper, Dai et al. introduced a novel under-sampling algorithm named Multi-Granularity Relabeled Under-Sampling (MGRU) (Dai et al., 2022). This technique enhances traditional under-sampling by focusing on potentially overlapping instances in local subspaces of features, thereby aiming to improve classification accuracy. The core advancement in MGRU lies in its ability to detect and remove majority class instances that contribute to class overlap, using a combination of Mahalanobis distance and local subspace analysis. This approach not only reduces class overlap but also preserves the essential structure of the data, which is often compromised in typical under-sampling methods.

Yu et al. proposed a method that tackles class imbalance by dynamically adjusting the weights of individual instances based on their learning difficulties, rather than adjusting class weights uniformly. It measures the difficulty of each instance by observing how quickly it is learned, similar to human learning processes. This approach uses a re-sampling algorithm that tracks prediction changes for each instance across training epochs to adjust their sampling probabilities. More weight is given to slower-learning instances to address the underrepresentation of minority classes and challenging instances in majority classes. The strategy has shown superior performance on class-imbalanced datasets and is supported by theoretical proofs of correctness and convergence. Although no sample is removed in this approach, the sample can be included in the instance selection category as the importance weights are assigned to the samples of both classes (S. Yu et al., 2022).

Farshidvar et al. proposed a two-phase clustering-based under-sampling method which focuses on retaining the general data pattern and distribution by utilizing a convex-hull-based clustering technique. In the first phase, the majority class is divided into clusters such that no minority class samples fall within the convex-hull of these clusters (Farshidvar et al., 2023). This is achieved by ensuring that each cluster's size is controlled, aiming to prevent any alteration in data distribution. In the second phase, the method employs ensemble learning where multiple classifiers are trained on the balanced datasets generated from the clusters. This approach enhances the robustness and predictive power of the model. The ensemble learning phase mitigates information loss by incorporating diverse datasets derived from different clusters.

Soltanzadeh et al. employed a metaheuristic optimization algorithm to address both class imbalance and class overlap in imbalanced datasets. Initially, the method involves separating input data into majority and minority classes and then generating an initial population of majority class subsets. These subsets are used to create under-sampled datasets that are assessed using a classifier to evaluate their effectiveness based on an evaluation metric. If the datasets are not optimal, the process iteratively generates new solutions through a metaheuristic algorithm until the termination condition, often a maximum number of generations is met. The versatility of the approach is highlighted by its adaptability to various metaheuristic algorithms, such as the Artificial Bee Colony (ABC), Particle Swarm Optimization (PSO), and Genetic Algorithm (GA), among which ABC shows superior performance in experiments (Soltanzadeh et al., 2023).

A novel Clustering-Based Noisy-Sample-Removed Under-sampling Scheme (NUS) is proposed in (H. Zhu et al., 2024) which addresses the challenges of imbalanced classification, which is prevalent in industries such as credit card fraud detection. The method starts by clustering the majority class samples and designating the furthest cluster center to define a hypersphere. Within this setup, it examines whether minority class samples fall inside or outside this hypersphere, identifying and removing those outside as noisy. A similar noise removal is applied to the majority class. NUS then employs under-sampling,

reducing the number of majority class samples to balance the dataset effectively. This integration of noise removal and under-sampling refines the dataset by focusing on high-quality, relevant samples for training classifiers.

Under-sampling Based on Minority Class Density (UBMD), developed by Sun et al., targets class imbalance by emphasizing the density characteristics of minority class samples (Z. Sun et al., 2024). This innovative method employs kernel density estimation to meticulously model the probability density distribution of the minority class. By doing so, it facilitates a focused filtering process that identifies and eliminates majority class samples residing within high-density zones of the minority class, thereby mitigating typical information loss associated with traditional under-sampling techniques. Additionally, UBMD introduces a concept called "sampling fitness", which assesses the value of each majority class sample in terms of richness and relevance, ensuring that only the most informative samples are retained. This method enhances classification tasks by maintaining the natural distribution characteristics of the majority class, effectively preserving critical information and improving accuracy.

### • Hybrid methods

According to some papers, a combination of under-sampling techniques and SMOTE performs better than using SMOTE solely. The combination of SMOTE and Tomek Links is presented in 2004, which generates synthetic samples using the SMOTE algorithm and then eliminates redundant samples using Tomek Links (Batista et al., 2004). SMOTE + Tomek Links reduces class overlap, leading to improved accuracy and reduced overfitting compared to SMOTE alone. This method tends to perform well on imbalanced datasets where the majority class contains noisy or redundant samples. SMOTE-ENN is another method that incorporates SMOTE and Wilson's ENN (Wilson, 1972). In this method, the SMOTE algorithm generated synthetic samples in the first step. Next, the samples misclassified by their three nearest neighbors are eliminated from the training set (Batista et al., 2004). SMOTE-ENN has demonstrated substantial improvements in identifying relevant samples without sacrificing accuracy, due to its ability to refine synthetic samples and clean up noisy data. It is particularly effective in high-dimensional datasets, where noise is a significant issue.

Agglomerative Hierarchical Clustering (AHC) is a hybrid method that uses clustering to balance datasets. In this method, the  $k$ -means algorithm is used to under-sample the majority class samples, and agglomerative hierarchical clustering is used to over-sample the minority class samples. AHC gathers the clusters from all levels of the resulting dendrograms and interpolates their centroids with the original minority samples (Cohen et al., 2006). SPIDER2 is another hybrid method that defines two types of samples, noisy and safe ones. Those samples which their  $k$ -nearest neighbors correctly classify them stay in a safe category; otherwise, they are grouped as noisy. Therefore, safe samples are easy to classify; meanwhile, noisy ones are hard to be labeled; thus, they need special attention. SPIDER employs a cleaning method on the majority class samples; then, it balances the distribution by over-sampling the minority borderline samples (Stefanowski & Wilk, 2008).

The authors in (Seiffert et al., 2009) combined two sampling methods: a base technique and the second method to complete the sampling procedure. They take advantage of three over-sampling techniques (ROS, SMOTE (Chawla et al., 2002), and borderline-SMOTE (H. Han et al., 2005)) with two under-sampling techniques (random under-sampling and Wilson's editing (Wilson, 1972)). Moreover, using the geometric mean criterion reveals that the classifier performs better when over-sampling techniques are applied before under-sampling methods in most cases. UnderOverBagging or UnderBagging to OverBagging method (S. Wang & Yao, 2009) considers combining the over-sampling and under-sampling processes. This method applies a resampling rate 'a%' in each iteration that is ranged from 10 % to 100 % and is

the multiple of 10. Hence, training the first classifiers is done with a lower number of samples than the last ones. A combination of synthetic minority over-sampling technique (SMOTE) and an under-sampling based on complementary neural network (CMTNN) is proposed in (Jeatrakul et al., 2010). This method tends to enhance the classification accuracy toward the rare class by using a combination of over-sampling and under-sampling methods. Also, it uses three types of learning algorithms, including ANN, SVM, and k-NN.

Ivotes is a bagging-based approach that integrates a rule-based ensemble with the SPIDER data level technique. This approach is robust to atypical data distributions in the minority classes and can automatically find the optimal number of bags (Błaszczyński et al., 2010). S-RSM (Sampling-response surface methodologies) proposed in (Tong et al., 2011) tends to determine the best resampling strategy by designing experiments (DOE) and response surface methodologies (RSM). This method applies the over-sampling and under-sampling methods simultaneously to find an adequate number of samples to eliminate from the majority class and duplicate in the minority class. In (Y. Liu et al., 2011), the authors first showed that the SVM classifier has weaknesses in dealing with imbalanced datasets. Then they proposed an ensemble method using SVM classifier and over-sampling and under-sampling techniques. The proposed method first over-samples the minority class to provide supplement any information for the training data and then uses under-sampling to avoid overfitting.

Another hybrid technique is proposed in (Ramentol et al., 2012) that combines SMOTE as an over-sampling technique and an under-sampling technique designed based on rough set theory. Wang proposed a hybrid method in (Q. Wang, 2014), which first trains SVM using the imbalanced training data and defines the hyperplane; afterward, it removes the majority samples that are far away from the hyperplane, assuming they have less information. To do over-sampling, it divides the new train set into  $k$  subsets and uses SMOTE to generate new samples in randomly selected subsets. Similarity-based under-sampling and normal distribution based over-sampling method (SUNDO) is presented in (Cateni et al., 2014). This method is a hybrid approach, including over-sampling and under-sampling phases. In the over-sampling phase, the goal is to find the places where it is more probable to generate samples there and avoid creating new samples near frequent samples. Also, the under-sampling phase performs based on detecting irregular data samples. An extension of SMOTE called Iterative-Partitioning Filter (IPF) is suggested in (Sáez et al., 2015). IPF filter is added to the SMOTE algorithm to control borderline and noisy samples produced after applying SMOTE and makes the boundary of classes more consistent. SCUT is a hybrid method for multi-class imbalanced data sets that does over-sampling for minority classes first and then applies a cluster-based under-sampling to the majority classes (Agrawal et al., 2015). By doing so, both between class and within-class imbalance are controlled.

In (Charte, Rivera, del Jesus, et al., 2015), the imbalance issue in multi-label data sets is investigated, and one random over-sampling and one random under-sampling technique are presented to balance the data set. DBSM is another hybrid algorithm consisting of a popular DBSCAN based under-sampling method and SMOTE algorithm (Sanguanmak & Hanskunat, 2016). EPRENNID is an evolutionary prototype reduction for nearest neighbor classifying of imbalanced data sets. In this method, both prototype generation and selection are employed to reduce overfitting (Vluymans et al., 2016). Mao et al. aimed at designing a hybrid method for sequential imbalanced data set in (Mao et al., 2017). Their main idea was to preserve the initial data's sequential distribution property after being over-sampled and under-sampled. In their proposed method, over-sampling and under-sampling are done based on the majority and minority classes' confidence regions built using the principle curve. In another method (F. Hu et al., 2019), neighborhood density is considered to decide whether under-sampling or over-sampling should be used in such a way that if the density of majority samples in an area is high, under-sampling is applied. Suppose the area has a balanced number of majority and minority samples. In that case, SMOTE is used,

and when the density of minority samples is high in that area, over-sampling is employed.

In Ref. (Susan & Kumar, 2019), a hybrid three-step preprocessing method is proposed in which first, an under-sampling method is applied on the majority class. Afterward, an over-sampling technique is used to expand the minority set, and finally, the under-sampling method is applied again, but this time to the new minority class. The under-sampling technique they used is called subspace optimization, designed based on particle swarm optimization algorithm. For over-sampling, SMOTE and three other SMOTE-based algorithms are chosen. Their intelligent combination of under-sampling and over-sampling methods could improve the learning performance in the end. A hybrid approach based on the density concept and clustering called Clustering and Density-Based Hybrid (CDBH) is proposed in (Mirzaei et al., 2021). In the over-sampling process, CDBH first clusters the minority class samples using the  $k$ -means algorithm, and then it obtains the densities of samples in each cluster. To generate the new minority samples, the denser minority samples will be chosen more likely by the roulette wheel selection operator. In the under-sampling process, the denser majority samples will have more chances to be chosen as the majority class representative, like the previous stage. In this method,  $IR$  will be one at the end.

Authors in (Mirzaei et al., 2022) introduced a hybrid sampling method called the hybrid sharing-based sampling technique (HSST) for classifying imbalanced data. This method applies a criterion called sharing score in both under-sampling and over-sampling stages to determine the importance of each sample in the feature space. In the over-sampling stage of the HSST method, the synthetic samples for the minority class are generated by interpolating between the more sparsely distributed samples. To under-sample the majority class, the denser samples from the majority class are selected to be removed. The binary tournament selection operator is used in both stages to perform the respective sampling processes based on probabilities. This method ensures  $IR$  becomes one at the end. RN-SMOTE, which stands for Reduced Noise SMOTE, is an enhanced hybrid technique that combines the Synthetic Minority Over-sampling Technique (SMOTE) with a noise reduction process using DBSCAN, a density-based clustering algorithm (Arafa et al., 2022). The method unfolds in three critical steps: initially, it oversamples the minority class using SMOTE to counteract the imbalance; then, it employs DBSCAN to pinpoint and eliminate noisy instances from the newly generated samples, addressing the common issue of noise introduction by SMOTE; finally, SMOTE is reapplied to the cleaned data to ensure the dataset remains balanced. This refined process not only preserves the diversity of the synthetic samples but also significantly enhances the quality of the training data by reducing noise, thereby improving the performance of various classifiers.

Dixit et al. introduced SMOTE-TLNN-DEPSO, a novel filtering-based over-sampling method for addressing issues in imbalanced classification due to noisy and borderline examples (Dixit & Mani, 2023). Traditional SMOTE and its variations often struggle with these problems due to their sensitivity to noise and difficulty in defining class boundaries. SMOTE-TLNN-DEPSO integrates SMOTE for initial synthetic sample generation, then applies a two-layer natural neighbors' technique (TLNN) for error detection without the need for parameter tuning. Instead of removing noisy and borderline examples, it employs DEPSO (Differential Evolution and Particle Swarm Optimization), to iteratively optimize and correct their attributes. This method not only maintains the imbalance ratio but also improves the decision boundary, making it particularly effective for datasets with significant noise in class attributes.

SMOTENN, introduced by Vairetti et al., is a novel hybrid resampling method for handling imbalanced big data classification (Vairetti et al., 2024). This method integrates both over-sampling and under-sampling techniques using a MapReduce framework, improving efficiency by performing both processes in a single pass. Specifically, SMOTENN combines SMOTE for over-sampling and Edited Nearest Neighbors

(ENN) for intelligent under-sampling. Key advantages of SMOTENN include its ability to preserve the quality of data while balancing the dataset effectively. It defines a neighborhood for each minority class sample, where it simultaneously generates synthetic samples and removes noisy majority class samples. This integrated approach not only addresses the imbalance but also enhances the overall prediction accuracy.

There are some papers in imbalanced data set problems that analyze data-level methods. For example, authors in (Napierala et al., 2010) present an experimental study on the impact of several factors in resampling methods in the imbalanced domain. Firstly, they reveal that the degradation in the classifier's performance is highly related to the number of borderline samples. In the situations that overlapping regions are large enough and at least 30 % of minority class samples are in the borderline's area, focused resampling techniques have superiority over random and cluster-based over-sampling techniques. On the contrary, while the number of minority samples in borderline areas is small, random and cluster-based over-sampling methods sufficiently improve the rare class samples' recognition.

In (Luengo et al., 2011), the effects of preprocessing techniques in an imbalanced domain have been analyzed. To do so, three sampling methods, including SMOTE (Chawla et al., 2002), SMOTE-ENN (Batista et al., 2004) and EUSCHC (S. García & Herrera, 2009) have been chosen. The paper reveals that the imbalanced ratio is not enough to measure the data complexity. Finally, the authors present two precise and straightforward rules to demonstrate suitable and unsuitable datasets for using the noted three methods. They consider intervals of data complexity measures' values for every dataset in which C4.5 and PART perform good or bad. Accordingly, the first rule is that good behavior for C4.5 and PART is achieved by an average high test AUC in the interval without overfitting. On the other hand, the second rule expresses that bad behavior occurs when there is overfitting and/or average test AUC is low in the interval. Also, they reveal that employing Fisher's Discriminant Ratio for measures of overlaps in feature values from different classes leads to achieving satisfactory results.

The work in (Kovács, 2019a) conducts a comprehensive comparison of 85 different over-sampling methods applied to 104 datasets with imbalanced class distributions. The aim of this work is to establish a new baseline in the field, identify the over-sampling principles that lead to the best results across general circumstances, and also provide guidance to practitioners on which over-sampling techniques are most suitable for use with particular types of datasets. Analyzing the performance of the different operating principles of over-sampling techniques can provide valuable insights that can guide and accelerate the development of new and improved over-sampling methods, by identifying the most successful principles.

## 5. Benchmarks, Software, and toolboxes

In this section, the frequently used data repositories are presented, and then some popular software and toolboxes are introduced.

### 5.1. Imbalanced benchmark data sets

In literature, the most widely used imbalanced data sets come from the KEEL repository (<http://www.keel.es/>). There are 145 data sets, including binary class data sets, multi-class data sets, some synthetic data sets with noisy and borderline samples (Napierala et al., 2010), and some preprocessed data sets. Also, the partitioned data sets using the 5-folds cross-validation procedure are available on this website. The imbalance ratio ranges from 1.8 to 100.4 for binary class and from 1.5 to 853 for multi-class data sets. To use the partitioned data sets efficiently, we have built a MATLAB data file including all binary class and multi-class data sets and provided a code to read from this file at <https://github.com/Bahar-nkr/KEEL-imbalanced-data-sets>.

Another well-known repository used for imbalanced learning

experiments is the UCI Machine Learning Repository (<https://archive.ics.uci.edu/>). There are 476 data sets in this repository, but they are not specifically imbalanced. Most of the time, either some classes of each data set are merged and considered as one class, or some classes are selected, and the samples of other classes are ignored so that the final data set is imbalanced.

Also, an imbalanced data set presented on the Kaggle website is the Credit Card data set, and the goal is to detect fraud using that. This data set includes credit card transactions in September 2013 in Europe (Dal Pozzolo et al., 2014) and is highly imbalanced with an imbalance ratio of 0.0017. Also, the number of samples and features in this data set is 284,807 and 28, respectively. Kaggle's website has other imbalanced data sets, including Hmeq and Promotio as well. Other imbalanced data sets are used in literature, such as DNA microarray data (H. Yu et al., 2013); however, they are not widespread and broadly used in imbalanced learning.

### 5.2. Software and toolboxes

In addition to the data repository, KEEL has an open-source software tool with the same name containing many algorithms for different data mining objectives (Triguero et al., 2017). One of the sections of this software is dedicated to imbalanced learning, and there exist several popular data-level algorithms that can be easily used. A complete set of statistical procedures are provided in this software, including parametric and nonparametric tests, to compare the algorithms pairwise (S. García et al., 2010). WEKA (Waikato environment for knowledge analysis) is another data mining software that includes some preprocessing techniques in the imbalanced domain as well (<https://www.cs.waikato.ac.nz/ml/weka/>). Both KEEL and WEKA are java-based and free, open-source software. Moreover, they are multi-platform, which means they can be used in Windows, Mac, and Linux and have a graphical user interface.

In python, a toolbox called Imbalanced-learn has been presented, offering several sampling methods commonly used in the imbalanced learning field (Lemaître et al., 2017). To boost applications and advancements in the field of imbalanced learning, the package smote-variants provides a python implementation of 85 over-sampling methods (Kovács, 2019b). This package provides multi-class compatible with 61 of the implemented binary over-sampling methods. It also includes a model selection framework that makes it easy to find the suitable over-sampling method for a specific dataset. Moreover, a variety of cross-validation and evaluation functionalities are provided to make using the package easier. The examples, documentation, and source code are available at <https://github.com/gykovacs/sMOTE-variants/>.

For R programming language, several packagers, such as ROSE (Rose et al., 2015), DMwR (Torgo & Torgo, 2013), unbalanced (Andrea et al., 2015), and CRAN, are created that include most of the practical and popular data-level methods. Moreover, many functions are publicly available in the MATLAB environment simulating data level approaches in the imbalance field. A software named Multi-Imbalance was recently developed in 2018 for multi-class imbalanced data classification (C. Zhang, Bi, et al., 2019). This software is developed in MATLAB and contains several preprocessing methods for the multi-class category of imbalanced data sets.

## 6. Performance evaluation

Evaluation criteria are essential for assessing predictive model performance, guiding the design and selection of machine learning algorithms. Typically, the process involves training a model, testing it on unseen data, and comparing predictions with actual outcomes. For classification problems, this means comparing predicted and true labels. These criteria also assist in selecting suitable model types and preprocessing methods, enabling researchers to experiment with various



models and identify the most effective one based on measured results (Y. Sun et al., 2009).

### 6.1. Taxonomy of evaluation metrics

Standard evaluation criteria, like classification accuracy, are widely used to assess predictive models and are effective for many applications. However, their reliability depends on the assumptions they make about the problem and what is considered important. Choosing the right metric requires aligning with project goals and stakeholder priorities. This selection becomes particularly challenging in imbalanced class distributions, where standard criteria may become unreliable or misleading, especially with extreme imbalance ratios (H. He & Ma, 2013). Standard evaluation criteria can produce sub-optimal classification models and misleading results when applied to imbalanced data, as they lack sensitivity to the challenges posed by imbalanced domains (Branco et al., 2015).

Imbalanced data classification often requires specialized evaluation criteria that prioritize the minority class, unlike standard criteria that treat all classes equally. Selecting appropriate metrics is challenging due to the limited examples from the minority class, which can hinder the training of effective models (Branco et al., 2015).

In 2018, Brzezinski et al. introduced a specialized visualization approach using a barycentric coordinate system and a 3D tetrahedron to analyze evaluation measures comprehensively. They adapted this technique for imbalanced data and proposed key properties to consider when evaluating classification performance. Their method highlights potential issues, such as inappropriate parameter settings for the  $F_\beta$ -score that can favor the majority class in imbalanced tasks. An online tool accompanies this approach, enabling users to analyze predefined and custom measures effectively (Brzezinski et al., 2018).

In another paper, Brzezinski et al. emphasize the importance of considering class proportions when interpreting evaluation measures. For instance, the  $F_1$ -score, which represents the harmonic mean of precision and recall, can be misleading in imbalanced datasets. An  $F_1$ -score of 0.7 may indicate good performance in balanced data but becomes much harder to achieve in cases of severe imbalance. This highlights the need for careful selection and interpretation of metrics that accurately reflect model performance across different imbalance ratios (Brzezinski et al., 2020).

In 2021, Stapor et al. critically analyzed evaluation methodologies for machine learning classifiers, identifying flaws in protocols and statistical methods that often produce unreliable results. They pointed out issues like improper cross-validation, biased metrics, and misapplied statistical tests. To improve evaluation reliability, they advocated for balanced metrics like the  $F_1$ -score for imbalanced data, robust cross-validation techniques, representative datasets, and non-parametric statistical tests like the Wilcoxon signed-rank test. They also emphasized transparency in reporting results and proposed independent evaluation platforms to ensure unbiased assessments, ultimately enhancing the credibility of classifier evaluations (Stapor et al., 2021).

Evaluation criteria for predictive models can be broadly categorized into two types: threshold metrics and ranking metrics, with a wide variety of measures available (Ferri et al., 2009; H. He & Ma, 2013). This categorization highlights the diversity of metrics that need to be considered to accurately assess model performance under different circumstances.

**A. Threshold metrics:** Threshold criteria are criteria that quantify the amount of classification prediction error. These criteria determine the fraction, ratio, or rate of times that predictions differ from the expected values (Ferri et al., 2009). Accuracy can be considered as the most widely used threshold criterion for classification problems:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Number of Total Predictions}} \quad (2)$$

While classification accuracy is commonly used, it is unsuitable for imbalanced data, as even a no-skill model can achieve high accuracy by predicting only the majority class. Most threshold metrics are derived from the confusion matrix, which provides detailed insights into a model's performance, including correctly or incorrectly predicted classes and error types. The confusion matrix for a two classes classification problem (binary) is presented in Fig. 2.

This does not mean that the metrics are limited for binary classification; it is just an easy way to understand what is being measured quickly. The confusion matrix returns four values, including true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), which are explained in the following:

- o TP: The number of positive samples that are classified as a positive class correctly.
- o TN: The number of negative samples that are classified as a negative class correctly.
- o FP: The number of negative samples that are classified as a positive class incorrectly.
- o FN: The number of positive samples that are classified as a negative class incorrectly.

Further, other criteria can be obtained from the confusion matrix to measure the performance of the classifier in both positive and negative classes: There are two groups of metrics (sensitivity-specificity and precision-recall) that focus on one class; thus, they may be useful for imbalanced classification problems (Branco et al., 2015).

#### ° Sensitivity-Specificity Metrics:

- **Sensitivity:** It is also called TPR (True Positive Rate), which indicates the classifier's ability to identify positive class instances correctly with a value in range 0 to 1, and is defined as:  $\frac{TP}{TP+FN}$
- **Specificity:** It is also called Selectivity or TNR (True Negative Rate), which indicates the classifier's ability to identify negative class instances correctly with a value in range 0 to 1, and is defined as:  $\frac{TN}{TN+FP}$

In imbalanced classification, sensitivity is often more informative than specificity. Combining both, the geometric mean (G-Mean) is a widely used metric that evaluates a classifier's ability to balance performance across minority and majority classes by correlating sensitivity and specificity, ensuring balanced accuracy (Branco et al., 2015). In this way, G-mean's low value indicates that the classifier is highly biased towards one of the classes, and a high value of G-mean means the classifier is predicting two classes well enough. G-mean is defined as follows:

$$G - \text{mean} = \sqrt[2]{\text{Sensitivity} \times \text{Specificity}} \quad (3)$$

#### ° Precision-Recall Metrics:

- **Precision:** It is also called PPV (Positive Predictive Value), which indicates the classifier's ability to avoid misclassifying negative class instances as the positive class, and is defined as:  $\frac{TP}{TP+FP}$
- **Recall:** It is also called TPR (True Positive Rate), which indicates the classifier's ability to identify positive class instances correctly with a value in range 0 to 1, and is defined as:  $\frac{TP}{TP+FN}$ . Recall has the same calculation as sensitivity.

It is also possible to combine precision and recall and create a widespread criterion that considers both concerns simultaneously, called the  $F_1$ -score.  $F_1$ -score takes advantage of a harmonic mean between precision and recall (Hazim Obaid et al., 2024). The  $F_1$ -score, as the harmonic mean of precision and recall, ensures both are reasonably high. Precision measures the percentage of correct positive predictions,

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Fig. 2. Confusion matrix for a two-class problem.

while recall reflects the proportion of actual positive cases identified. A low  $F_1$ -score indicates an imbalance between precision and recall, where one is sacrificed for the other (Branco et al., 2015). Correspondingly, a high  $F_1$ -score will be achieved when both precision and recall are high:

$$F_1 - \text{Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

There is an abstraction of the  $F_1$ -score called the  $F_\beta$ -score, where the balance of precision and recall is controlled by a coefficient called Beta:

$$F_\beta - \text{score} = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}} \quad (5)$$

Ease of understanding and implementation can be considered as the essential advantages of the threshold metrics. The limitation of these metrics is that they assume the class distribution observed in the training dataset matches the test set's distribution. Although this is often the case, the performance can be quite misleading if not (H. He & Ma, 2013).

**B. Ranking Metrics:** Ranking metrics, the second category of evaluation criteria, are crucial for scenarios where accurately separating classes is highly important (Ferri et al., 2009). These metrics are suitable for classifiers that predict class membership scores or probabilities. Different thresholds can be applied to evaluate performance, with models that score well across a broad range of thresholds demonstrating good class separation and thus being ranked higher (Fernández, García, Galar, et al., 2018b). ROC curve and PR curves are known as the most popular ranking metrics.

- **ROC Curve:** ROC (Receiver Operating Characteristic) summarizes a binary classifier's ability to distinguish classes by plotting the true positive rate versus the false positive rate at various threshold values. Each point on the curve represents the ratio for a specific threshold. The X-axis represents the false positive rate, and the Y-axis represents the true positive rate. A no-skill classifier is shown as a diagonal line, with any point below it performing worse than an unskilled classifier. An ideal classifier is represented as a point at the top left (Ferri et al., 2009) (Fig. 3).

The ROC curve is used to assess a single classifier, but comparing multiple classifiers is challenging unless one consistently outperforms the other. Hence, the area under the ROC curve (AUC) is used to measure performance, with an AUC of 0.5 indicating no skill and 1.0 representing a perfect classifier (Ferri et al., 2009).

- **PR Curve:** The Precision-Recall (PR) curve is an alternative to the ROC curve, focusing on the minority class. It plots precision (y-axis) and recall (x-axis) for different thresholds, with each point representing the precision-to-recall ratio. An unskilled classifier is shown as a horizontal line, with precision proportional to the minority class instances, while a perfect classifier is at the top right of the graph (Ferri et al., 2009) (Fig. 4).

Similar to the ROC curve, the PR curve evaluates a single classifier's performance, but comparing multiple classifiers can be difficult.

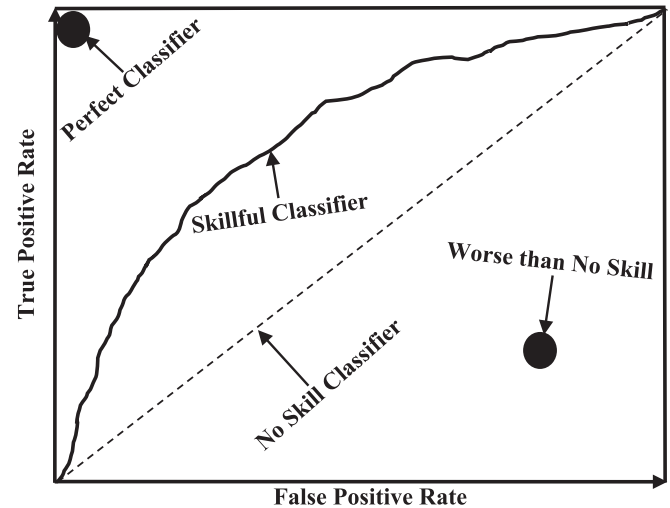


Fig. 3. The ROC Curve of Perfect, Skillful, No Skill, and Worse than No Skill Classifiers.

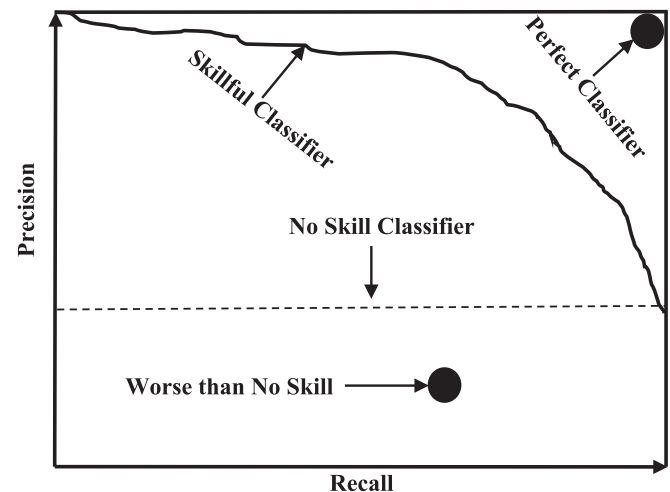


Fig. 4. The PR Curve of Perfect, Skillful, No Skill, and Worse than No Skill Classifiers.

Therefore, the PR AUC (Area Under the Curve) is used to compare classifiers, especially in imbalanced classification problems, as it focuses on the minority class.

Other metrics are less widely used, such as probabilistic metrics (Ferri et al., 2009) and visual-based metrics (Brzezinski et al., 2018). Selecting the right evaluation criteria is crucial for designing classification models (Ferri et al., 2009). With many criteria available, choosing the right one depends on the problem type, classification purpose, and class importance. Fig. 5 summarizes the proposed criteria and their applications.



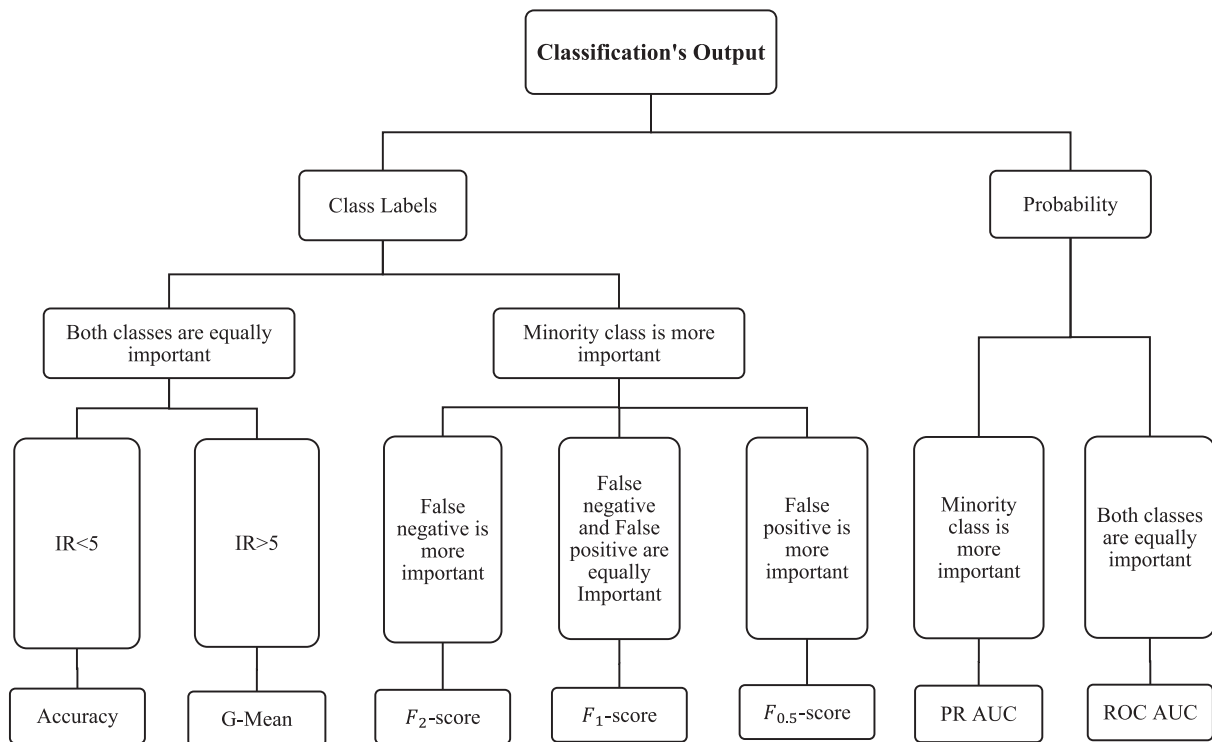


Fig. 5. Guide to choosing the appropriate criteria for evaluating imbalanced data classifiers.

## 7. Challenges and future trends

In recent years, several papers such as (V. García et al., 2012; López et al., 2013; Orriols-Puig & Bernadó-Mansilla, 2009) present imbalance class distribution as the most obvious factor of significant performance loss in imbalanced classification problems. For binary class problems, an imbalance ratio, which presents the amounts of inequality between classes, is defined as the proportion of the majority class size to the size of the minority class. This kind of inequality is mainly known as the between-class imbalance, and it can be in the range of 1:10,000 (H. He & Shen, 2007; Kubat et al., 1998).

Although experimental studies demonstrate that the inequality between classes' distribution is the main factor in deteriorating classification performance, other factors affect classifier's modeling in detecting rare events (Batista et al., 2004; Haixiang et al., 2017; López et al., 2013; Y. Sun et al., 2009). Other practical facts, including small disjuncts, lack of density, noise, dataset shift, class overlapping, imperfect data, imbalanced big data, and imbalanced image classification, influence classification's performance and need to be considered in the imbalanced domains to get better results. These factors are explained below:

- **Small Disjuncts:** in classification problems, there is a situation where a class consists of several subclasses or subclusters. This within-class imbalance happens when samples of a class are gathered from different subconcepts (Jo & Japkowicz, 2004). These subconcepts, known as the small disjuncts, do not always have the same number of examples (Datta et al., 2017; Napierala & Stefanowski, 2016; Prati et al., 2004b; Ramyachitra & Manikandan, 2014; Weiss, 2010). Small disjuncts problems deteriorate the classification procedure since within-class subconcepts are implicit in most cases and increase the learning concept's complexity (Y. Sun et al., 2009). To deal with the small disjuncts problem in sampling methods, identifying these subclasses in the dataset by taking advantage of clustering methods before tending to equal the class distribution could achieve better results (Douzas et al., 2018). Also, various solutions

have been proposed in the literature (Fernández, García, Galar, et al., 2018b) to reduce the problem of small disjuncts, the most important of which are as follows:

- Obtaining more training data:** Those poorly represented classes may create small disjuncts, especially in the minority class. By achieving the new data in under-represented regions, the problem can be reduced.
- Use adequate inductive bias:** To avoid introducing the artificial small disjuncts created by a mismatch between the algorithm bias and the data at hand, richer inductive bias can solve the problem.
- Using more appropriate metrics:** Selecting appropriate metrics can identify disjuncts correctly in the presence of noisy samples or when class overlapping exists. Another possible solution is employing different metrics independent of class imbalance.
- Better control pruning:** If pruning parameters are determined correctly, an appropriate trade-off between disjunct sizes and classification performance can be established. By selecting a strong pruning technique, most small disjuncts will be removed by generalizing the classification rules. That is while without considering pruning, the likelihood of smaller disjuncts will be highly increased.
- Using ensembles:** Up to now, various ensemble algorithms have been proposed to solve the problem of class imbalance. Since ensembles are based on integrating classifiers' output, small disjuncts' disadvantages may be averaged out by applying ensembles.
- **Lack of Density:** the lack of density or insufficiency of information in the training data is one of the most common challenges in classification when the dataset size is not big enough (Krawczyk, 2016; Raudys & Jain, 1991; Stefanowski, 2015). The lack of density causes the induction algorithms not to have enough information about the data distribution to generalize its model. However, it becomes a more severe problem when the data is high dimensional and imbalanced (Wasikowski & Chen, 2010). In the imbalanced data classification with a fixed imbalanced ratio, the classifier's performance is directly related to the dataset's size, i.e., as the size of the dataset increases, the classifier's performance becomes better. However, if the dataset's size is not large enough, the classifier may

not generalize data characteristics. Moreover, the classifier could overfit the training data, with an undesirable performance in testing samples (Raudys & Jain, 1991). Increasing dataset size means increasing information, and more information helps the algorithm build its model more accurately, so it helps the classifier detect rare events better (Japkowicz & Stephen, 2002).

- **Noise:** in the class imbalanced problems, noise has a more significant impact on rare classes, i.e., as the minority class contains fewer samples, less noisy patterns are needed to impact the classifier's performance over them (Krawczyk, 2016; Rivera, 2017; Weiss, 2004). Furthermore, minority class examples may be identified as noise by the learning algorithm, but noise could also be considered minority class examples since both are rare in the dataset (Beyan & Fisher, 2015). There are two kinds of noise in the literature: a feature (or attribute) and class noise. Class noise is generally supposed to be more detrimental than attribute noise in machine learning (Frénay & Verleysen, 2014). Feature noise affects the observed values of features, whereas class noise changes the observed class values (e.g., changing the label of a minority class sample to the majority class label). An experimental study in (Khoshgoftaar et al., 2011) shows the benefits of bagging methods without replacement in such cases and recommends applying noise reduction techniques before boosting approaches. Also, in Ref. (L. Ma & Fan, 2017), authors have utilized clustering to identify noise samples before over-sampling.
- **Dataset Shift:** the situation where training and test samples pursue different distributions is known as the data set shift problem (Alaiz-Rodríguez & Japkowicz, 2008; Fernández et al., 2011; Several, 2010). Dataset shift can be divided into three types (Fernández, García, Herrera, et al., 2018):
  - a. **Prior Probability Shift:** It happens when the training and test sets have a different class distribution. This problem can be directly investigated by applying a stratified cross-validation scheme, such that both sets include the same number of samples per class.
  - b. **Covariate Shift:** It happens when the input attribute values' distributions are different among the training and test sets. This problem mainly occurs when the data is partitioned for validation purposes. The stratified  $k$ -fold cross-validation is the most used procedure for this task causing this kind of induced dataset shift as it shuffles the samples between the different folds randomly.
  - c. **Concept Shift:** It happens when the relationship between the input and class variables changes. This problem, which is usually known as "Concept Drift" in specialized literature, represents the most formidable challenge between the different kinds of dataset shift.

Dataset shift can influence all kinds of classification problems, and it appears because of sampling selection bias issues. Dataset shift is widespread in imbalanced domain issues due to the unequal number of samples in different classes. In datasets with a high value of the imbalanced ratio, due to the low number of examples in minority class, it is more sensitive to the singular classification errors (Batuwita & Palade, 2013).

- **Class Overlapping:** although the main goal in the imbalanced domain is to induce the learning algorithm to separate rare events from the relevant ones correctly, it would not be challenging if classes are discriminative enough from each other. However, several papers reveal that the class overlapping diminishes the classification performance more than class imbalance (Batista et al., 2005; V. García et al., 2006, 2007; Prati et al., 2004a; Wilk et al., 2016; Xiong et al., 2010). Moreover, based on the conducted experiments by researchers, using the learning algorithm C4.5, this claim that when the class overlapping is more extensive, a class imbalance has a more substantial influence in degrading the performance of the induced classifier is supported. Thus, different class overlapping levels in some feature spaces make it hard for the classifier to induce its discriminative rules. The most popular metric used to calculate the

degree of overlap for a given dataset is the maximum Fisher's discriminant ratio, known as  $F_1$ -measure, or simply  $F_1$  too (Ho & Basu, 2002). Note that it should not be confused with the  $F_1$ -score performance metric. This metric can be obtained for every individual feature (one dimension) as follows:

$$F_1 - Measure = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 - \sigma_2^2} \quad (6)$$

where  $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$  are the means and variances of the two classes, respectively. Lastly, the maximum value for all features is considered as  $F_1$ . A small value for the  $F_1$  metric means that the dataset has a high degree of overlapping. To deal with the issue mentioned above, several authors have suggested taking advantage of sampling techniques in the pre-processing step. Conversely, many papers have revealed that sampling might not be enough to solve class overlapping problems. The studies have pointed out that as the class overlapping happens due to irrelevant and redundant features, feature selection techniques could be an excellent way to deal with this issue. The relationship between the overlapping and class imbalance problems and the performance and complexity of learned models has been investigated in (Denil & Trapenberg, 2010).

- **Imperfect Data:** one of the prevalent problems in real-world applications is imperfect data. Some problems related to data imperfection contain incompleteness, imprecision, inconsistency, and uncertainty. These problems occur for many reasons, including inadequate data transcription, faulty sensors, data collection errors, unreliable data acquisition or transmission sources, and the lack of data representation standards (Fernández, García, Galar, et al., 2018b). When dealing with imperfect and imbalanced data, possible disadvantages are two-fold: Firstly, the methods presented to confront imbalanced data may reduce their performance because of the bad quality of data. Secondly, the performance of methods to deal with imperfect data may be influenced by imbalanced data. Traditionally, class imbalance and data imperfection are treated distinctly. However, there are few attempts with imbalance and imperfection jointly. For example, Ref. (Sowah et al., 2016) presents data clustering integrated with under-sampling to identify spurious data points.
- **Imbalanced Big Data:** due to generating massive amounts of information by modern systems, practical solutions must be developed for processing them computationally effective. The class imbalance affects big data, which causes increasing the challenge in the learning systems. Increasing the data volume can lead to prohibiting the existing methods, and the nature of the problem can lead to additional difficulties. Big imbalanced data can be created by different specific areas such as social networks or computer vision, which forces us to work with specific kinds of data, including graphs and tensors or video sequences. As a result, in addition to being scalable and efficient, the algorithms should handle heterogeneous and atypical data (Krawczyk, 2016). Some research works investigate the relationship between classification complexity and class imbalance. For example, the authors in (Weng & Poon, 2006) employ data complexity measures to obtain insights about a data level technique's behavior according to data projection in two text classification tasks. Also, Luengo et al. (Luengo et al., 2011) investigated the relationship between the Fisher discriminant ratio with the class imbalance ratio and the classifiers' performance with and without treatment for class imbalance.
- **Imbalanced Image Classification:** an emerging and important area of research within imbalanced data classification is the imbalanced image classification. This area presents unique challenges due to the nature of image data and the limitations of traditional over-sampling techniques like SMOTE when applied to images. For example, SMOTE-based approaches are often ill-suited for handling image

data, because they generate synthetic samples by interpolating between existing ones, which can lead to unrealistic and noisy images that do not align well with the original data distribution (Buda et al., 2018; Dablain et al., 2023; Douzas & Bacao, 2018).

Recent research has begun to address these challenges using advanced techniques such as Generative Adversarial Networks (GANs). GAN-based methods have shown promise in generating realistic synthetic images that can help balance the class distribution without introducing significant noise (Douzas & Bacao, 2018; Goodfellow et al., 2014; Mariani et al., 2018). The work by Buda et al. (Buda et al., 2018) highlights the effectiveness of such techniques in improving the performance of classifiers on imbalanced image datasets.

Moreover, other innovative approaches include data augmentation techniques tailored for image data and the use of deep learning models that inherently handle class imbalance by learning robust feature representations. These methods are crucial as they open new avenues for effectively tackling the imbalanced image classification problem, which is increasingly prevalent in fields such as medical imaging, autonomous driving, and facial recognition (Esteva et al., 2017; Shorten & Khoshgoftaar, 2019).

By considering the problems above, it can be concluded that although between-class imbalance appears to be the most problematic issue, these factors should not be overseen because sometimes imbalance is not the main problem. For example, data sets in which classes have small disjuncts confront the problem with classic data-level methods such as SMOTE as they may generate samples in the other classes' regions. Hence, future researchers should pay more attention to such problems rather than just balancing the data or finding redundant and useful samples. It is clear that data sets are different from different aspects, such as the distribution of classes or the way samples are located in each class and the sample number of classes; therefore, the pre-processing method for each data set should be different.

It appears that by clustering, the locality of each sample can be considered more, which provides more information for data generation or reduction. That is why methods are tending to be cluster-based and partly-inclusive. However, most of the clustering algorithms have some parameters to tune. The number of clusters should be predefined, so clustering techniques that overcome this problem can be helpful to have improvement. Another issue to point out is that recently, the deep learning field has become very prevalent in most fields, and it has been shown that imbalanced data has a negative effect on such methods. As a result, it should be considered, and new methods should be designed to be employed in this field (Buda et al., 2018). The field of deep learning has made significant strides in addressing the challenges associated with imbalanced data. Innovative techniques such as GANs for synthetic data generation (Mariani et al., 2018), modified loss functions (T. Y. Lin et al., 2020), and cost-sensitive learning (S. H. Khan et al., 2018) have been particularly impactful. These advancements have enhanced the capability of deep learning models to handle class imbalance, leading to improved performance and robustness in various applications. Future research should continue to explore and refine these methods.

Moreover, the application of data augmentation in imbalanced learning remains an area of active research. Recent literature, such as (Chen et al., 2024; Johnson & Khoshgoftaar, 2019; A. A. Khan et al., 2024) provide detailed explorations of the latest augmentation methods, highlighting their effectiveness in improving deep learning performance on imbalanced datasets. These studies underscore the growing importance of data augmentation in imbalanced learning and point to it as a promising area for future research.

Today, most data samples belong to more than one label. Such data sets are called multi-label, and class imbalance happens in them as well. It can be another future research field as it is an important issue, and there is a minimal amount of research on it.

In our discussion of challenges and future trends in imbalanced data classification, several issues overlap with those identified by Krawczyk

(Krawczyk, 2016). Both works identify the impact of lack of density, noise, class overlapping, and imbalanced big data. However, our work also highlights additional challenges such as small disjuncts, dataset shift, imperfect data, imbalanced image classification, and imbalanced problems in the deep learning field. Recognizing these additional challenges is crucial for advancing the field of imbalanced data classification and addressing the complexities of modern data environments.

Since Krawczyk's 2016 paper, the challenges associated with imbalanced image classification and imbalanced problems in the deep learning field have gained significant attention. While Krawczyk identified critical challenges in imbalanced data learning, the increased focus on deep learning and image classification in recent years highlights the growing importance and unique challenges of these domains.

## 8. Conclusion

In imbalanced data sets, samples are unequally distributed among the classes, which causes problems for classical classifiers. There are many methods proposed to deal with this problem, a group of which is data level approaches. In this paper, we aimed at reviewing these methods. We first presented some critical applications of imbalanced learning to highlight the essence of working in this field and showed how prevalent this problem is in the real world. Then, we proposed several categorizations from different perspectives for data-level methods, including balancing perspective, interaction with the learning algorithm perspective, class perspective, inclusiveness perspective, resampling in ensembles perspective, and deep learning methods. We elaborated each category in detail afterward. Subsequently, we reviewed the recently proposed and state-of-the-art data-level methods from a balancing perspective as it is the most common categorization in the literature. Also, we presented the related data sets and toolboxes and the evaluation criteria that are essential for researchers for doing experiments. In the end, we discussed the present challenges, which makes dealing with imbalanced data sets even more difficult, and future trends for researchers to pursue.

## Declaration of competing interest

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

## Data availability

No data was used for the research described in the article.

## References

- Abd El-Naby, A., Hemdan, E.-E.-D., & El-Sayed, A. (2023). An efficient fraud detection framework with credit card imbalanced data in financial services. *Multimedia Tools and Applications*, 82(3), 4139–4160.
- Abd Elrahman, S. M., & Abraham, A. (2013). A review of class imbalance problem. *Journal of network and innovative. Computing*, 1, 332–340. [www.mirlabs.net/jnic/index.html](http://www.mirlabs.net/jnic/index.html).
- Abdi, L., & Hashemi, S. (2016). To combat multi-class imbalanced problems by means of over-sampling techniques. *IEEE Transactions on Knowledge and Data Engineering*, 28(1), 238–251. <https://doi.org/10.1109/TKDE.2015.2458858>
- Abokadr, S., Azman, A., Hamdan, H., & Amelina, N. (2023). Handling Imbalanced Data for Improved Classification Performance: Methods and Challenges. *2023 3rd International Conference on Emerging Smart Technologies and Applications, ESmarTA 2023*, 1–8. <https://doi.org/10.1109/eSmarTA59349.2023.10293442>.
- Achouch, M., Dimitrova, M., Ziane, K., Sattarpanah Karganroudi, S., Dhoubi, R., Ibrahim, H., & Adda, M. (2022). On predictive maintenance in industry 4.0: Overview, Models, and Challenges. *Applied Sciences (Switzerland)*, 12(16), 8081. <https://doi.org/10.3390/app12168081>
- Agrawal, A., Viktor, H. L., & Paquet, E. (2015). SCUT: Multi-class imbalanced data classification using SMOTE and cluster-based undersampling. *IC3K 2015 - Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 1, 226–234. <https://doi.org/10.5220/0005595502260234>.



- Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022). Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare*, 10(3), 541.
- Al Banna, M. H., Taher, K. A., Kaiser, M. S., Mahmud, M., Rahman, M. S., Hosen, A. S. M. S., & Cho, G. H. (2020). Application of artificial intelligence in predicting earthquakes: State-of-the-art and future challenges. *IEEE Access*, 8, 192880–192923.
- Alaiz-Rodríguez, R., & Japkowicz, N. (2008). Assessing the impact of changing environments on classifier performance. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: Vol. 5032 LNAI (pp. 13–24). [https://doi.org/10.1007/978-3-540-68825-9\\_2](https://doi.org/10.1007/978-3-540-68825-9_2)
- Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqan, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial fraud detection based on machine learning: A systematic literature review. *Applied Sciences*, 12(19), 9637.
- Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with class imbalance problem: A review. *International Journal of Advances in Soft Computing and Its Applications*, 7(3), 176–204.
- Aljeldah, M. M. (2022). Antimicrobial resistance and its spread is a global threat. *Antibiotics*, 11(8), 1082.
- Anand, A., Pugalenth, G., Fogel, G. B., & Suganthan, P. N. (2010). An approach for classification of highly imbalanced data using weighting and undersampling. *Amino Acids*, 39(5), 1385–1391. <https://doi.org/10.1007/s00726-010-0595-2>
- Anderson, B., & Adey, P. (2012). Governing events and life: “Emergency” in UK civil contingencies. *Political Geography*, 31(1), 24–33. <https://doi.org/10.1016/j.polgeo.2011.09.002>
- Ando, S., & Huang, C. Y. (2017). Deep over-sampling framework for classifying imbalanced data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: Vol. 10534 LNAI (pp. 770–785). [https://doi.org/10.1007/978-3-319-71249-9\\_46](https://doi.org/10.1007/978-3-319-71249-9_46)
- Andrea, A., Pozzolo, D., Caelen, O., Bontempi, G., Andrea, M., & Pozzolo, D. (2015). Package ‘unbalanced’.
- Araba, A., El-Fishawy, N., Badawy, M., & Radad, M. (2022). RN-SMOTE: Reduced noise SMOTE based on DBSCAN for enhancing imbalanced data classification. *Journal of King Saud University - Computer and Information Sciences*, 34(8), 5059–5074. <https://doi.org/10.1016/j.jksuci.2022.06.005>
- Asnari, Maulidevi, N. U., & Surendro, K. (2022). SMOTE-LOF for noise identification in imbalanced data classification. *Journal of King Saud University - Computer and Information Sciences*, 34(6), 3413–3423. <https://doi.org/10.1016/j.jksuci.2021.01.014>
- Azaria, A., Richardson, A., Kraus, S., & Subrahmanian, V. S. (2014). Behavioral analysis of insider threat: A survey and bootstrapped prediction in imbalanced data. *IEEE Transactions on Computational Social Systems*, 1(2), 135–155. <https://doi.org/10.1109/TCSS.2014.2377811>
- Barandela, R., Sánchez, J. S., & Valdivinos, R. M. (2003). New applications of ensembles of classifiers. *Pattern Analysis and Applications*, 6(3), 245–256. <https://doi.org/10.1007/s10044-003-0192-z>
- Barandela, R., Valdivinos, R. M., Salvador Sánchez, J., Ferri, F. J., Sánchez, J. S., & Ferri, F. J. (2004). *The Imbalanced Training Sample Problem: Under or over Sampling?* In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 3138, pp. 806–814). [https://doi.org/10.1007/978-3-540-27868-9\\_88](https://doi.org/10.1007/978-3-540-27868-9_88)
- Barella, V. H., Costa, E. P., & Carvalho, A. C. P. L. F. (2014). ClusterOSS : A new undersampling method for imbalanced learning. *Brazilian Conference on Intelligent Systems*, 1–6.
- Barua, S., Islam, M. M., & Murase, K. (2011). A novel synthetic minority oversampling technique for imbalanced data set learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: Vol. 7063 LNCS (Issue PART 2, pp. 735–744). [https://doi.org/10.1007/978-3-642-24958-7\\_85](https://doi.org/10.1007/978-3-642-24958-7_85)
- Barua, S., Islam, M. M., Yao, X., & Murase, K. (2014). MWMOTE - majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(2), 405–425. <https://doi.org/10.1109/TKDE.2012.232>
- Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20–29. <https://doi.org/10.1145/1007730.1007735>
- Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2005). Balancing strategies and class overlapping. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: Vol. 3646 LNCS (pp. 24–35). [https://doi.org/10.1007/11552253\\_3](https://doi.org/10.1007/11552253_3)
- Batuwita, R., & Palade, V. (2013). Class imbalance learning methods for support vector machines. *Imbalanced Learning: Foundations, Algorithms, and Applications*, 83–99. <https://doi.org/10.1002/9781118646106.ch5>
- Bekkar, M., & Alitouch, T. A. (2013). Imbalanced data Learning approaches review. *International Journal of Data Mining & Knowledge Management Process*, 3(4), 15–33. <https://doi.org/10.5121/ijdkp.2013.3402>
- Beyan, C., & Fisher, R. (2015). Classifying imbalanced data sets using similarity based hierarchical decomposition. *Pattern Recognition*, 48(5), 1653–1672. <https://doi.org/10.1016/j.patcog.2014.10.032>
- Bhatta, S., & Dang, J. (2023). Seismic damage prediction of RC buildings using machine learning. *Earthquake Engineering & Structural Dynamics*, 52(11), 3504–3527.
- Błaszczyszński, J., Deckert, M., Stefanowski, J., & Wilk, S. (2010). Integrating selective pre-processing of imbalanced data with votes ensemble. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [https://doi.org/10.1007/978-3-642-13529-3\\_17](https://doi.org/10.1007/978-3-642-13529-3_17)
- Branco, P., Torgo, L., & Ribeiro, R. (2015). A Survey of Predictive Modelling under Imbalanced Distributions.
- Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, 49(2), 1–50. <https://doi.org/10.1145/2907070>
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140. <https://doi.org/10.1007/bf00058655>
- Brzezinski, D., Stefanowski, J., Susmaga, R., & Szczech, I. (2020). On the dynamics of classification measures for imbalanced and streaming data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8), 2868–2878. <https://doi.org/10.1109/TNNLS.2019.2899061>
- Brzezinski, D., Stefanowski, J., Susmaga, R., & Szczech, I. (2018). Visual-based analysis of classification measures and their properties for class imbalanced problems. *Information Sciences*, 462, 242–261. <https://doi.org/10.1016/j.ins.2018.06.020>
- Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249–259. <https://doi.org/10.1016/j.neunet.2018.07.011>
- Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [https://doi.org/10.1007/978-3-642-01307-2\\_43](https://doi.org/10.1007/978-3-642-01307-2_43)
- Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2012). DBSMOTE: Density-based synthetic minority over-sampling technique. *Applied Intelligence*, 36(3), 664–684. <https://doi.org/10.1007/s10489-011-0287-y>
- Cao, Q., & Wang, S. (2011). Applying over-sampling technique based on data density and cost-sensitive SVM to imbalanced learning. *Proceedings - 2011 4th International Conference on Information Management, Innovation Management and Industrial Engineering, ICIMI 2011*, 2, 543–548. <https://doi.org/10.1109/ICIMI.2011.276>
- Castellanos, F. J., Valero-Mas, J. J., Calvo-Zaragoza, J., & Rico-Juan, J. R. (2018). Oversampling imbalanced data in the string space. *Pattern Recognition Letters*, 103, 32–38. <https://doi.org/10.1016/j.patrec.2018.01.003>
- Cateni, S., Colla, V., & Vannucci, M. (2014). A method for resampling imbalanced datasets in binary classification tasks for real-world problems. *Neurocomputing*, 135, 32–41. <https://doi.org/10.1016/j.neucom.2013.05.059>
- Charte, F., Rivera, A. J., del Jesus, M. J., & Herrera, F. (2015). Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, 163, 3–16. <https://doi.org/10.1016/j.neucom.2014.08.091>
- Charte, F., Rivera, A. J., del Jesus, M. J., & Herrera, F. (2015). MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, 89, 385–397. <https://doi.org/10.1016/j.knsys.2015.07.019>
- Chawla, N. V. (2009). Data Mining for Imbalanced Datasets: An overview. *Data Mining and Knowledge Discovery Handbook*, 30(1), 875–886. [https://doi.org/10.1007/978-0-387-09823-4\\_45](https://doi.org/10.1007/978-0-387-09823-4_45)
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 2838, 107–119. [https://doi.org/10.1007/978-3-540-39804-2\\_12](https://doi.org/10.1007/978-3-540-39804-2_12)
- Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: Latest research, applications and future directions. In *Artificial Intelligence Review (Vol. 57, Issue 6)*. <https://doi.org/10.1007/s10462-024-10759-6>
- Cohen, G., Hilarío, M., Sax, H., Hugonnet, S., & Geissbühler, A. (2006). Learning from imbalanced data in surveillance of nosocomial infection. *Artificial Intelligence in Medicine*, 37(1), 7–18. <https://doi.org/10.1016/j.artmed.2005.03.002>
- D’Addabbo, A., & Maglietta, R. (2015). Parallel selective sampling method for imbalanced and large data classification. *Pattern Recognition Letters*, 62, 61–67. <https://doi.org/10.1016/j.patrec.2015.05.008>
- Dablain, D., Krawczyk, B., & Chawla, N. V. (2023). DeepSMOTE: Fusing deep Learning and SMOTE for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9), 6390–6404. <https://doi.org/10.1109/TNNLS.2021.3136503>
- Dai, Q., Liu, J. wei, & Liu, Y. (2022). Multi-granularity relabeled under-sampling algorithm for imbalanced data. *Applied Soft Computing*, 124, 109083. <https://doi.org/10.1016/j.asoc.2022.109083>
- Dal Pozzolo, A., Caelen, O., Le Borgne, Y. A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41(10), 4915–4928. <https://doi.org/10.1016/j.eswa.2014.02.026>
- Dalzochio, J., Kunst, R., Pignaton, E., Binotto, A., Sanyal, S., Favilla, J., & Barbosa, J. (2020). Machine learning and reasoning for predictive maintenance in industry 4.0: Current status and challenges. *Computers in Industry*, 123, Article 103298.
- Datta, S., Nag, S., Mullick, S. S., & Das, S. (2017). Diversifying Support Vector Machines for Boosting using Kernel Perturbation: Applications to Class Imbalance and Small Disjuncts. <http://arxiv.org/abs/1712.08493>
- de Haro-García, A., & García-Pedrajas, N. (2011). A scalable method for instance selection for class-imbalance datasets. *2011 11th International Conference on Intelligent Systems Design and Applications*, 1383–1390. <https://doi.org/10.1109/ISDA.2011.6121853>
- Denil, M., & Trappenberg, T. (2010). Overlap versus imbalance. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: Vol. 6085 LNAI (pp. 220–231). [https://doi.org/10.1007/978-3-642-13059-5\\_22](https://doi.org/10.1007/978-3-642-13059-5_22)

- Devaraj, A., Murthy, D., & Dontula, A. (2020). Machine-learning methods for identifying social media-based requests for urgent help during hurricanes. *International Journal of Disaster Risk Reduction*, 51, Article 101757.
- Ding, H., Huang, N., & Cui, X. (2024). Leveraging GANs data augmentation for imbalanced medical image classification. *Applied Soft Computing*, 165(July), Article 112050. <https://doi.org/10.1016/j.asoc.2024.112050>
- Dixit, A., & Mani, A. (2023). Sampling technique for noisy and borderline examples problem in imbalanced classification. *Applied Soft Computing*, 142, Article 110361. <https://doi.org/10.1016/j.asoc.2023.110361>
- Douzas, G., & Bacao, F. (2017). Self-organizing map oversampling (SOMO) for imbalanced data set learning. *Expert Systems with Applications*, 82, 40–52. <https://doi.org/10.1016/j.eswa.2017.03.073>
- Douzas, G., & Bacao, F. (2018). Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications*, 91, 464–471. <https://doi.org/10.1016/j.eswa.2017.09.030>
- Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences*, 465, 1–20. <https://doi.org/10.1016/j.ins.2018.06.056>
- Escobar Díaz Guerrero, R., Carvalho, L., Bocklitz, T., Popp, J., & Oliveira, J. L. (2024). A Data Augmentation Methodology to Reduce the Class Imbalance in Histopathology Images. *Journal of Imaging Informatics in Medicine*, 37(4), 1767–1782. <https://doi.org/10.1007/s10278-024-01018-9>
- Ester, M., Kriegl, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 96(34), 226–231.
- Esteve, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. <https://doi.org/10.1038/nature21056>
- Farooq, Z., Rocklöv, J., Wallin, J., Abiri, N., Sewe, M. O., Sjödin, H., & Semenza, J. C. (2022). Artificial intelligence to predict West Nile virus outbreaks with eco-climatic drivers. *The Lancet Regional Health-Europe*, 17.
- Farshidvard, A., Hooshmand, F., & MirHassani, S. A. (2023). A novel two-phase clustering-based under-sampling method for imbalanced classification problems. *Expert Systems with Applications*, 213, Article 119003. <https://doi.org/10.1016/j.eswa.2022.119003>
- Farsi, M., Daneshkhah, A., Far, A. H., Chatrabgoun, O., & Montasari, R. (2018). Crime data mining, threat analysis and prediction. *Cyber Criminology*, 183–202.
- Feng, W., Dauphin, G., Huang, W., Quan, Y., Bao, W., Wu, M., & Li, Q. (2019). Dynamic synthetic minority over-sampling technique-based rotation forest for the classification of imbalanced hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7), 2159–2169. <https://doi.org/10.1109/JSTARS.2019.2922297>
- Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018a). Cost-Sensitive Learning. In *Learning from Imbalanced Data Sets* (pp. 63–78). Springer International Publishing. [https://doi.org/10.1007/978-3-319-98074-4\\_4](https://doi.org/10.1007/978-3-319-98074-4_4)
- Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018b). Learning from Imbalanced Data Sets. In *Learning from Imbalanced Data Sets*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-98074-4>
- Fernández, A., García, S., & Herrera, F. (2011). Addressing the classification with imbalanced data: Open problems and new challenges on class distribution. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6678 LNAI(PART 1), 1–10. [https://doi.org/10.1007/978-3-642-21219-2\\_1](https://doi.org/10.1007/978-3-642-21219-2_1)
- Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from imbalanced data: Progress and challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 61, 863–905. <https://doi.org/10.1613/jair.11192>
- Ferri, C., Hernández-Orallo, J., & Modroui, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1), 27–38. <https://doi.org/10.1016/j.patrec.2008.08.010>
- Frénay, B., & Verleysen, M. (2014). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5), 845–869. <https://doi.org/10.1109/TNNLS.2013.2292894>
- Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). 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. <https://doi.org/10.1109/TSMCC.2011.2161285>
- Galar, M., Fernández, A., Barrenechea, E., & Herrera, F. (2013). EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recognition*, 46(12), 3460–3471. <https://doi.org/10.1016/j.patcog.2013.05.006>
- García-Pedrajas, N. (2011). Evolutionary computation for training set selection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(6), 512–523. <https://doi.org/10.1002/widm.44>
- García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, 180(10), 2044–2064. <https://doi.org/10.1016/j.ins.2009.12.010>
- García, S., & Herrera, F. (2009). Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. *Evolutionary Computation*, 17(3), 275–306. <https://doi.org/10.1162/evco.2009.17.3.275>
- García, V., Alejo, R., Sánchez, J. S., Sotoca, J. M., & Mollineda, R. A. (2006). Combined effects of class imbalance and class overlap on instance-based classification. In *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: Vol. 4224 LNCS (pp. 371–378). [https://doi.org/10.1007/11875581\\_45](https://doi.org/10.1007/11875581_45)
- García, V., Mollineda, R. A., Sánchez, J. S., Alejo, R., & Sotoca, J. M. (2007). When overlapping unexpectedly alters the class imbalance effects, 2, 499–506. [https://doi.org/10.1007/978-3-540-72849-8\\_63](https://doi.org/10.1007/978-3-540-72849-8_63)
- García, V., Sánchez, J. S., & Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25(1), 13–21. <https://doi.org/10.1016/j.knsys.2011.06.013>
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 3(January), 2672–2680. [https://doi.org/10.1007/978-3-658-40442-0\\_9](https://doi.org/10.1007/978-3-658-40442-0_9)
- Grabec, I., Švegl, E., & Sok, M. (2019). A method for automatic medical diagnosis. *Statistics, Optimization and Information Computing*, 7(1), 26–39. <https://doi.org/10.19139/soic.v7i1.414>
- Gu, J., Zhao, L., Yue, X., Arshad, N. I., & Mohamad, U. H. (2023). Multistage quality control in manufacturing process using blockchain with machine learning technique. *Information Processing & Management*, 60(4), Article 103341.
- Guo, H., & Viktor, H. L. (2004). Learning from imbalanced data sets with boosting and data generation. *ACM SIGKDD Explorations Newsletter*, 6(1), 30–39. <https://doi.org/10.1145/1007730.1007736>
- Guo, S., Liu, Y., Chen, R., Sun, X., & Wang, X. (2019). Improved SMOTE algorithm to deal with imbalanced activity classes in Smart homes. *Neural Processing Letters*, 50(2), 1503–1526. <https://doi.org/10.1007/s11063-018-9940-3>
- Gupta, A., & Katarya, R. (2020). Social media based surveillance systems for healthcare using machine learning: A systematic review. *Journal of Biomedical Informatics*, 108, Article 103500.
- Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73, 220–239. <https://doi.org/10.1016/j.eswa.2016.12.035>
- Halpern, N. A. (2018). Early warning systems for hospitalized pediatric patients. *JAMA - Journal of the American Medical Association*, 319(10), 981–982. <https://doi.org/10.1001/jama.2018.1524>
- Han, H., Wang, W.-Y.-Y., & Mao, B.-H.-H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. *Lecture Notes in Computer Science*, 3644(PART 1), 878–887. [https://doi.org/10.1007/11538059\\_91](https://doi.org/10.1007/11538059_91)
- Han, W., Huang, Z., Li, S., & Jia, Y. (2019). Distribution-sensitive unbalanced data over-sampling method for Medical diagnosis. *Journal of Medical Systems*, 43(2), 39. <https://doi.org/10.1007/s10916-018-1154-8>
- Harlman, R., & Uchida, K. (2018). Data- and algorithm-hybrid approach for imbalanced data problems in deep neural network. *International Journal of Machine Learning and Computing*, 8(3), 208–213. <https://doi.org/10.18178/ijmlc.2018.8.3.689>
- Hart, P. E. (1968). The condensed Nearest neighbor rule. *IEEE Transactions on Information Theory*, 14(3), 515–516. <https://doi.org/10.1109/TIT.1968.1054155>
- Hazim Obaid, Z., Mirzaei, B., & Darroudi, A. (2024). An efficient automatic modulation recognition using time-frequency information based on hybrid deep learning and bagging approach. *Knowledge and Information Systems*, 66(4), 2607–2624. <https://doi.org/10.1007/s10115-023-02041-y>
- He, H., Bai, Y., García, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the International Joint Conference on Neural Networks*, 1322–1328. <https://doi.org/10.1109/IJCNN.2008.4633969>
- He, H., & García, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- He, H., & Ma, Y. (2013). Imbalanced learning: Foundations, algorithms, and applications. In H. He & Y. Ma (Eds.), *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley. <https://doi.org/10.1002/9781118646106>
- He, H., & Shen, X. (2007). A ranked subspace learning method for gene expression data classification. *Proceedings of the 2007 International Conference on Artificial Intelligence, ICAI 2007*, 1, 358–364.
- He, K., Chen, X., Yu, X., Dong, C., & Zhao, D. (2024). Evaluation and prediction of compound geohazards in highly urbanized regions across China's Greater Bay Area. *Journal of Cleaner Production*, 449, Article 141641.
- Hegde, J., & Rokseth, B. (2020). Applications of machine learning methods for engineering risk assessment—a review. *Safety Science*, 122, Article 104492.
- Hensman, P., & Masko, D. (2015). *The impact of imbalanced training data for convolutional neural networks* (pp. 1313–1317). Isbi: PhD.
- Ho, T. K., & Basu, M. (2002). Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3), 289–300. <https://doi.org/10.1109/34.990132>
- Hoens, T. R., Qian, Q., Chawla, N. V., & Zhou, Z. H. (2012). Building decision trees for the multi-class imbalance problem. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: Vol. 7301 LNAI (Issue PART 1, pp. 122–134). [https://doi.org/10.1007/978-3-642-30217-6\\_11](https://doi.org/10.1007/978-3-642-30217-6_11)
- Hoque, A., Raj, J., Saha, A., & Bhattacharya, P. (2020). Earthquake magnitude prediction using machine learning technique. *Trends in Computational Intelligence, Security and Internet of Things: Third International Conference, ICCISIoT 2020, Tripura, India, December 29-30, 2020, Proceedings*, 3, 37–53.
- Hoyos-Osorio, J., Alvarez-Meza, A., Daza-Santacoloma, G., Orozco-Gutierrez, A., & Castellanos-Dominguez, G. (2021). Relevant information undersampling to support imbalanced data classification. *Neurocomputing*, 436, 136–146. <https://doi.org/10.1016/j.neucom.2021.01.033>
- Hu, F., Yu, C., Dai, J., & Liu, K. (2019). A mixed sampling method for imbalanced data based on neighborhood density. In *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analytics*. <https://doi.org/10.1109/ICCCBDA.2019.8725685>



- Hu, S., Liang, Y., Ma, L., & He, Y. (2009). MSMOTE: Improving classification performance when training data is imbalanced. *2nd International Workshop on Computer Science and Engineering, WCSE 2009*, 2, 13–17. <https://doi.org/10.1109/WCSE.2009.756>.
- Huang, D., Wang, S., & Liu, Z. (2021). A systematic review of prediction methods for emergency management. *International Journal of Disaster Risk Reduction*, 62, Article 102412.
- Huang, Z., Yang, C., Chen, X., Huang, K., & Xie, Y. (2020). Adaptive over-sampling method for classification with application to imbalanced datasets in aluminum electrolysis. *Neural Computing and Applications*, 32(11), 7183–7199. <https://doi.org/10.1007/s00521-019-04208-7>
- Huang, Z., Zheng, H., Li, C., & Che, C. (2024). Application of machine Learning-based K-means clustering for financial fraud detection. *Academic Journal of Science and Technology*, 10(1), 33–39.
- Ibrahim, M. S., Dong, W., & Yang, Q. (2020). Machine learning driven smart electric power systems: Current trends and new perspectives. *Applied Energy*, 272, Article 115237.
- Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5), 429–449. <https://doi.org/10.3233/ida-2002-6504>
- Jeatrakul, P., Wong, K. W., & Fung, C. C. (2010). Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: Vol. 6444 LNCS (Issue PART 2, pp. 152–159). [https://doi.org/10.1007/978-3-642-17534-3\\_19](https://doi.org/10.1007/978-3-642-17534-3_19).
- Jo, T., & Japkowicz, N. (2004). Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter*, 6(1), 40–49. <https://doi.org/10.1145/1007730.1007737>
- Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), 27. <https://doi.org/10.1186/s40537-019-0192-5>
- Jomthanachai, S., Wong, W.-P., & Lim, C.-P. (2021). An application of data development analysis and machine learning approach to risk management. *Ieee Access*, 9, 85978–85994.
- Kang, Q., Chen, X., Li, S., & Zhou, M. (2017). A noise-filtered under-sampling scheme for imbalanced classification. *IEEE Transactions on Cybernetics*, 47(12), 4263–4274. <https://doi.org/10.1109/TCYB.2016.2606104>
- Kashef, S., Nezamabadi-pour, H., & Nikpour, B. (2018). Multilabel feature selection: A comprehensive review and guiding experiments. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(2), e1240.
- Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys*, 52(4), 1–36. <https://doi.org/10.1145/3343440>
- Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalances problems: Combination, implementation and evaluation. *Expert Systems with Applications*, 244(May 2023). <https://doi.org/10.1016/j.eswa.2023.122778>.
- Khan, A. A., Zhang, T., Huang, X., & Usmani, A. (2023). Machine learning driven smart fire safety design of false ceiling and emergency response. *Process Safety and Environmental Protection*, 177, 1294–1306.
- Khan, S. H., Hayat, M., Bennamoun, M., Soheli, F. A., & Togneri, R. (2018). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8), 3573–3587. <https://doi.org/10.1109/TNNLS.2017.2732482>
- Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2011). Comparing boosting and bagging techniques with noisy and imbalanced data. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, 41(3), 552–568. <https://doi.org/10.1109/TSMCA.2010.2084081>
- Kim, H. J., Jo, N. O., & Shin, K. S. (2016). Optimization of cluster-based evolutionary undersampling for the artificial neural networks in corporate bankruptcy prediction. *Expert Systems with Applications*, 59, 226–234. <https://doi.org/10.1016/j.eswa.2016.04.027>
- Kim, S., Kim, H., & Namkoong, Y. (2016). Ordinal classification of imbalanced data with application in Emergency and Disaster information Services. *IEEE Intelligent Systems*, 31(5), 50–56. <https://doi.org/10.1109/MIS.2016.27>
- King, G., & Zeng, L. (2001). Logistic regression in Rare events data. *Political Analysis*, 9(2), 137–163. <https://doi.org/10.1093/oxfordjournals.pan.a004868>
- Kocher, G., & Kumar, G. (2021). Machine learning and deep learning methods for intrusion detection systems: Recent developments and challenges. *Soft Computing*, 25(15), 9731–9763.
- Kovács, G. (2019a). An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing Journal*, 83, Article 105662. <https://doi.org/10.1016/j.asoc.2019.105662>
- Kovács, G. (2019b). Smote-variants: A python implementation of 85 minority oversampling techniques. *Neurocomputing*, 366, 352–354. <https://doi.org/10.1016/j.neucom.2019.06.100>
- Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5(4), 221–232. <https://doi.org/10.1007/s13748-016-0094-0>
- Kubat, M. (2000). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *Fourteenth International Conference on Machine Learning*, 4(c), 2–6.
- Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30(2–3), 195–215. <https://doi.org/10.1023/a:1007452223027>
- Lai, J.-P., Chang, Y.-M., Chen, C.-H., & Pai, P.-F. (2020). A survey of machine learning models in renewable energy predictions. *Applied Sciences*, 10(17), 5975.
- Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2101(August), 63–66. [https://doi.org/10.1007/3-540-48229-6\\_9](https://doi.org/10.1007/3-540-48229-6_9)
- Lee, H., Park, M., & Kim, J. (2016). Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. In *Proceedings - International Conference on Image Processing*. <https://doi.org/10.1109/ICIP.2016.7533053>
- Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18, 1–5.
- Lin, J., Keogh, E., Fu, A., & Van Herle, H. (2005). Approximations to magic: Finding unusual medical time series. *18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*, 329–334.
- Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2), 318–327. <https://doi.org/10.1109/TPAMI.2018.2858826>
- Lin, W.-C.-C., Tsai, C.-F.-F., Hu, Y.-H.-H., & Jhang, J.-S.-S. (2017). Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409–410, 17–26. <https://doi.org/10.1016/j.ins.2017.05.008>
- Linardos, V., Drakaki, M., Tzionas, P., & Karnavas, Y. L. (2022). Machine learning in disaster management: Recent developments in methods and applications. *Machine Learning and Knowledge Extraction*, 4(2).
- Liu, C., & Hsieh, P. Y. (2020). Model-based synthetic sampling for imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 32(8), 1543–1556. <https://doi.org/10.1109/TKDE.2019.2905559>
- Liu, X. Y., Wu, J., & Zhou, Z. H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(2), 539–550. <https://doi.org/10.1109/TSMCB.2008.2007853>
- Liu, Y., Liu, Y., Yu, B. X. B., Zhong, S., & Hu, Z. (2023). Noise-robust oversampling for imbalanced data classification. *Pattern Recognition*, 133, Article 109008. <https://doi.org/10.1016/j.patcog.2022.109008>
- Liu, Y., Yu, X., Huang, J. X., & An, A. (2011). Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. *Information Processing and Management*, 47(4), 617–631. <https://doi.org/10.1016/j.ipm.2010.11.007>
- Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining review. *International Journal of Computer Science and Network (IJCSN)*, 2(1).
- López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250, 113–141. <https://doi.org/10.1016/j.ins.2013.07.007>
- López, V., Triguero, I., Carmona, C. J., García, S., & Herrera, F. (2014). Addressing imbalanced classification with instance generation techniques: IPAD-IDE. *Neurocomputing*, 126, 15–28. <https://doi.org/10.1016/j.neucom.2013.01.050>
- Lu, X., Ye, X., & Cheng, Y. (2024). An overlapping minimization-based over-sampling algorithm for binary imbalanced classification. *Engineering Applications of Artificial Intelligence*, 133, Article 108107. <https://doi.org/10.1016/j.engappai.2024.108107>
- Luengo, J., Fernández, A., García, S., & Herrera, F. (2011). Addressing data complexity for imbalanced data sets: Analysis of SMOTE-based oversampling and evolutionary undersampling. *Soft Computing*, 15(10), 1909–1936. <https://doi.org/10.1007/s00500-010-0625-8>
- Ma, J., Afolabi, D. O., Ren, J., & Zhen, A. (2019). Predicting seminal quality via imbalanced Learning with Evolutionary safe-level synthetic minority over-sampling technique. *Cognitive Computation*. <https://doi.org/10.1007/s12559-019-09657-9>
- Ma, L., & Fan, S. (2017). CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. *BMC Bioinformatics*, 18(1), 169. <https://doi.org/10.1186/s12859-017-1578-z>
- Maalouf, M., & Siddiqi, M. (2014). Weighted logistic regression for large-scale imbalanced and rare events data. *Knowledge-Based Systems*, 59, 142–148. <https://doi.org/10.1016/j.knsys.2014.01.012>
- Maciejewski, T., & Stefanowski, J. (2011). Local neighbourhood extension of SMOTE for mining imbalanced data. In *IEEE SSCI 2011: Symposium Series on Computational Intelligence - CIDM 2011: 2011 IEEE Symposium on Computational Intelligence and Data Mining* (pp. 104–111). <https://doi.org/10.1109/CIDM.2011.5949434>
- Malekloo, A., Ozer, E., AlHamaydeh, M., & Girolami, M. (2022). Machine learning and structural health monitoring overview with emerging technology and high-dimensional data source highlights. *Structural Health Monitoring*, 21(4), 1906–1955.
- Mani, I., & Zhang, I. (2003). kNN approach to unbalanced data distributions: A case study involving information extraction. *Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets*.
- Mao, W., Jiang, M., Wang, J., & Li, Y. (2017). Online extreme Learning machine with hybrid sampling strategy for sequential imbalanced data. *Cognitive Computation*, 9(6), 780–800. <https://doi.org/10.1007/s12559-017-9504-2>
- Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., & Malossi, C. (2018). BAKAN: Data Augmentation with Balancing GAN. *ArXiv Preprint ArXiv:1803.09655*. <http://arxiv.org/abs/1803.09655>
- Mathew, J., Pang, C. K., Luo, M., & Leong, W. H. (2018). Classification of imbalanced data by oversampling in kernel space of support vector machines. *IEEE Transactions on Neural Networks and Learning Systems*, 29(9), 4065–4076. <https://doi.org/10.1109/TNNLS.2017.2751612>
- Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1), 92–122. <https://doi.org/10.1007/s10618-012-0295-5>
- Mirzaei, B. (2024). A novel clustering-based over-sampling technique for imbalanced data sets. In *2024 32nd International Conference on Electrical Engineering (ICEE)* (pp. 1–7). <https://doi.org/10.1109/ICEE63041.2024.10668141>

- Mirzaei, B., Nikpour, B., & Nezamabadi-pour, H. (2021). CDBH: A clustering and density-based hybrid approach for imbalanced data classification. *Expert Systems with Applications*, 164, Article 114035. <https://doi.org/10.1016/j.eswa.2020.114035>
- Mirzaei, B., Nikpour, B., & Nezamabadi-Pour, H. (2020). An under-sampling technique for imbalanced data classification based on DBSCAN algorithm. *8th Iranian joint congress on fuzzy and intelligent systems. CFIS*, 2020, 21–26. <https://doi.org/10.1109/CFIS49607.2020.9238718>
- Mirzaei, B., Rahmati, F., & Nezamabadi-pour, H. (2022). A score-based preprocessing technique for class imbalance problems. *Pattern Analysis and Applications*, 25(4), 913–931.
- Napierala, K., & Stefanowski, J. (2016). Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, 46(3), 563–597. <https://doi.org/10.1007/s10844-015-0368-1>
- Napierala, K., Stefanowski, J., & Wilk, S. (2010). Learning from imbalanced data in presence of noisy and borderline examples. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [https://doi.org/10.1007/978-3-642-13529-3\\_18](https://doi.org/10.1007/978-3-642-13529-3_18)
- Nayak, T., Bhat, N., Bhat, V., Shetty, S., Javed, M., & Nagabhushan, P. (2019). Automatic segmentation and breast density estimation for cancer detection using an efficient watershed algorithm. In *Lecture Notes in Networks and Systems* (Vol. 43, pp. 347–358). [https://doi.org/10.1007/978-981-13-2514-4\\_29](https://doi.org/10.1007/978-981-13-2514-4_29)
- Nekooimehr, I., & Lai-Yuen, S. K. (2016). Adaptive semi-supervised weighted oversampling (A-SUWO) for imbalanced datasets. *Expert Systems with Applications*, 46, 405–416. <https://doi.org/10.1016/j.eswa.2015.10.031>
- Newaz, A., & Haq, F. S. (2022). A novel hybrid sampling framework for imbalanced Learning. *ArXiv Preprint. ArXiv:2208.09619*.
- Ng, W. W. Y., Hu, J., Yeung, D. S., Yin, S., & Roli, F. (2015). Diversified sensitivity-based undersampling for imbalance classification problems. *IEEE Transactions on Cybernetics*, 45(11), 2402–2412. <https://doi.org/10.1109/TCYB.2014.2372060>
- Nguyen, H. M., Cooper, E. W., & Kamei, K. (2011). Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1), 4. <https://doi.org/10.1504/ijkesdp.2011.039875>
- Nikolaou, M., Pavlopoulou, A., Georgakilas, A. G., & Kyrodimos, E. (2018). The challenge of drug resistance in cancer treatment: A current overview. *Clinical and Experimental Metastasis*, 35(4), 309–318. <https://doi.org/10.1007/s10585-018-9903-0>
- Nikpour, B., & Nezamabadi-pour, H. (2018). HTSS: A hyper-heuristic training set selection method for imbalanced data sets. *Iran Journal of Computer Science*, 1(2), 109–128. <https://doi.org/10.1007/s42044-018-0009-2>
- Nikpour, B., & Nezamabadi-pour, H. (2019). A memetic approach for training set selection in imbalanced data sets. *International Journal of Machine Learning and Cybernetics*, 10(11), 3043–3070. <https://doi.org/10.1007/s13042-019-01000-w>
- Nikpour, B., Shabani, M., & Nezamabadi-Pour, H. (2017). Proposing new method to improve gravitational fixed nearest neighbor algorithm for imbalanced data classification. In *2nd Conference on Swarm Intelligence and Evolutionary Computation*. <https://doi.org/10.1109/CSIEC.2017.7940167>
- Niu, G., Yang, P., Zheng, Y., Cai, X., & Qin, H. (2021). Automatic quality control of crowdsourced rainfall data with multiple noises: A machine learning approach. *Water Resources Research*, 57(11), Article e2020WR029121.
- Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F., & Ajayi, S. (2022). Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques. *Journal of Building Engineering*, 45, Article 103406.
- Orriols-Puig, A., & Bernadó-Mansilla, E. (2009). Evolutionary rule-based systems for imbalanced data sets. *Soft Computing*, 13(3), 213–225. <https://doi.org/10.1007/s00500-008-0319-7>
- Ouadah, A., Zemmouchi-Ghomari, L., & Salhi, N. (2022). Selecting an appropriate supervised machine learning algorithm for predictive maintenance. *The International Journal of Advanced Manufacturing Technology*, 119(7), 4277–4301.
- Pan, T., Pedrycz, W., Yang, J., & Wang, J. (2024). An improved generative adversarial network to oversample imbalanced datasets. *Engineering Applications of Artificial Intelligence*, 132, Article 107934.
- Pham, A.-D., Ngo, N.-T., Truong, T. T. H., Huynh, N.-T., & Truong, N.-S. (2020). Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. *Journal of Cleaner Production*, 260, Article 121082.
- Piri, S., Delen, D., & Liu, T. (2018). A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. *Decision Support Systems*, 106, 15–29. <https://doi.org/10.1016/j.dss.2017.11.006>
- Pourroostaei Ardakani, S., Liang, X., Mengistu, K. T., So, R. S., Wei, X., He, B., & Cheshmehzangi, A. (2023). Road car accident prediction using a machine-learning-enabled data analysis. *Sustainability*, 15(7), 5939.
- Prati, R. C., Batista, G. E. A. P. A., & Monard, M. C. (2004a). Class imbalances versus class overlapping: An analysis of a learning system behavior. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 2972, 312–321. [https://doi.org/10.1007/978-3-540-24694-7\\_32](https://doi.org/10.1007/978-3-540-24694-7_32)
- Prati, R. C., Batista, G. E. A. P. A., & Monard, M. C. (2004b). Learning with class skew and small disjuncts. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 3171, pp. 296–306). [https://doi.org/10.1007/978-3-540-28645-5\\_30](https://doi.org/10.1007/978-3-540-28645-5_30)
- Rahmati, F., Nezamabadi-pour, H., & Nikpour, B. (2020). A gravitational density-based mass sharing method for imbalanced data classification. *SN Applied Sciences*, 2(2), 260. <https://doi.org/10.1007/s42452-020-2039-2>
- Ramentol, E., Caballero, Y., Bello, R., & Herrera, F. (2012). SMOTE-rSB\*: A hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. *Knowledge and Information Systems*, 33(2), 245–265. <https://doi.org/10.1007/s10115-011-0465-6>
- Ramentol, E., Gondres, I., Lajes, S., Bello, R., Caballero, Y., Cornelis, C., & Herrera, F. (2016). Fuzzy-rough imbalanced learning for the diagnosis of high voltage circuit breaker maintenance: The SMOTE-FRST-2T algorithm. *Engineering Applications of Artificial Intelligence*, 48, 134–139. <https://doi.org/10.1016/j.engappai.2015.10.009>
- Ramyachitra, D., & Manikandan, P. (2014). Imbalanced dataset classification and solutions: A review. *International Journal of Computing and Business Research (IJCBR)*, 5(4), 2229–2166.
- Raudys, S. J., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for Practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3), 252–264. <https://doi.org/10.1109/34.75512>
- Rivera, W. A. (2017). Noise reduction a priori synthetic over-sampling for class imbalanced data sets. *Information Sciences*, 408, 146–161. <https://doi.org/10.1016/j.ins.2017.04.046>
- Rose, P., Lunardon, A. N., Menardi, G., Torelli, N., & Lunardon, M. N. (2015). *Lunardon, Menardi, Torelli - 2014 - Package ROSE*.
- Sáez, J. A., Krawczyk, B., & Woźniak, M. (2016). Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition*, 57, 164–178. <https://doi.org/10.1016/j.patcog.2016.03.012>
- Sáez, J. A., Luengo, J., Stefanowski, J., & Herrera, F. (2015). SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences*, 291(C), 184–203. <https://doi.org/10.1016/j.ins.2014.08.051>
- Sanguanmak, Y., & Hanskunatani, A. (2016). DBSM: The combination of DBSCAN and SMOTE for imbalanced data classification. In *2016 13th International Joint Conference on Computer Science and Software Engineering*. <https://doi.org/10.1109/JCSSE.2016.7748928>
- Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: An overview from machine learning perspective. *Journal of Big Data*, 7, 1–29.
- Saryzadi, S., Nikpour, B., & Nezamabadi-Pour, H. (2018). NPC: Neighbors' progressive competition algorithm for classification of imbalanced data sets. In *Proceedings - 3rd Iranian Conference on Signal Processing and Intelligent Systems*. <https://doi.org/10.1109/ICSPIS.2017.8311584>
- Schapiro, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5(2), 197–227. <https://doi.org/10.1007/bf00116037>
- Seiffert, C., Khoshgoftaar, T. M., & Van Hulse, J. (2009). Hybrid sampling for imbalanced data. *Integrated Computer-Aided Engineering*, 16(3), 193–210. <https://doi.org/10.3233/ICA-2009-0314>
- Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, 40(1), 185–197. <https://doi.org/10.1109/TSMCA.2009.2029559>
- Several. (2010). Dataset shift in machine Learning. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(1), 274. <https://doi.org/10.1111/j.1467-985X.2009.00624.10.x>
- Shah, V. (2021). Machine Learning algorithms for Cybersecurity: Detecting and preventing threats. *Revista Espanola de Documentacion Científica*, 15(4), 42–66.
- Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep Learning. *Journal of Big Data*, 6(1), 1–48. <https://doi.org/10.1186/s40537-019-0197-0>
- Soltanzadeh, P., Feizi-Derakhshi, M. R., & Hashemzadeh, M. (2023). Addressing the class-imbalance and class-overlap problems by a metaheuristic-based under-sampling approach. *Pattern Recognition*, 143, Article 109721. <https://doi.org/10.1016/j.patcog.2023.109721>
- Soltanzadeh, P., & Hashemzadeh, M. (2021). RCSMOTE: Range-controlled synthetic minority over-sampling technique for handling the class imbalance problem. *Information Sciences*, 542, 92–111. <https://doi.org/10.1016/j.ins.2020.07.014>
- Sowah, R. A., Agebure, M. A., Mills, G. A., Koumadi, K. M., & Fiawoo, S. Y. (2016). New cluster undersampling technique for class imbalance Learning. *International Journal of Machine Learning and Computing*, 6(3), 205–214. <https://doi.org/10.18178/ijmlc.2016.6.3.599>
- Staňková, K., Brown, J. S., Dalton, W. S., & Gatenby, R. A. (2019). Optimizing cancer treatment using game theory: A review. *JAMA Oncology*, 5(1), 96–103. <https://doi.org/10.1001/jamaoncol.2018.3395>
- Stapor, K., Ksieniewicz, P., García, S., & Woźniak, M. (2021). How to design the fair experimental classifier evaluation. *Applied Soft Computing*, 104. <https://doi.org/10.1016/j.asoc.2021.107219>
- Stefanowski, J. (2015). Dealing with data difficulty factors while learning from imbalanced data. In *Challenges in Computational Statistics and Data Mining* (Vol. 605, pp. 333–363). [https://doi.org/10.1007/978-3-319-18781-5\\_17](https://doi.org/10.1007/978-3-319-18781-5_17)
- Stefanowski, J., & Wilk, S. (2008). Selective pre-processing of imbalanced data for improving classification performance. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5182 LNCS, 283–292. [https://doi.org/10.1007/978-3-540-85836-2\\_27](https://doi.org/10.1007/978-3-540-85836-2_27)
- Sun, C., Yan, Z., Li, Q., Zheng, Y., Lu, X., & Cui, L. (2018). Abnormal group-based joint medical fraud detection. *IEEE Access*, 7, 13589–13596.
- Sun, J., Li, H., Fujita, H., Fu, B., & Ai, W. (2020). Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. *Information Fusion*, 54, 128–144. <https://doi.org/10.1016/j.inffus.2019.07.006>
- Sun, Y., Kamel, M. S., Wong, A. K. C., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12), 3358–3378. <https://doi.org/10.1016/j.patcog.2007.04.009>



- Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4), 687–719. <https://doi.org/10.1142/S0218001409007326>
- Sun, Z., Ying, W., Zhang, W., & Gong, S. (2024). Undersampling method based on minority class density for imbalanced data. *Expert Systems with Applications*, 249, Article 123328. <https://doi.org/10.1016/j.eswa.2024.123328>
- Sundarkumar, G. G., & Ravi, V. (2015). A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. *Engineering Applications of Artificial Intelligence*, 37, 368–377. <https://doi.org/10.1016/j.engappai.2014.09.019>
- Susan, S., & Kumar, A. (2019). SSO maj -SMOTE-SSO min : Three-step intelligent pruning of majority and minority samples for learning from imbalanced datasets. *Applied Soft Computing Journal*, 78, 141–149. <https://doi.org/10.1016/j.asoc.2019.02.028>
- Tahir, M. A., Kittler, J., & Yan, F. (2012). Inverse random under sampling for class imbalance problem and its application to multi-label classification. *Pattern Recognition*, 45(10), 3738–3750. <https://doi.org/10.1016/j.patcog.2012.03.014>
- Tang, S., & Chen, S. P. (2008). The generation mechanism of synthetic minority class examples. In *5th Int. Conference on Information Technology and Applications in Biomedicine, ITAB 2008 in Conjunction with 2nd Int. Symposium and Summer School on Biomedical and Health Engineering*. <https://doi.org/10.1109/ITAB.2008.4570642>
- Tang, Y., Zhang, Y. Q., & Chawla, N. V. (2009). SVMs modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(1), 281–288. <https://doi.org/10.1109/TSMCB.2008.2002909>
- Tao, L., Li, H., Wang, F., Liu, M., Tang, Z., & Wang, Q. (2024). An Adaptive safe-region diversity oversampling algorithm for imbalanced classification. *IEEE Access*, 12, 63713–63724. <https://doi.org/10.1109/ACCESS.2024.3396155>
- Tao, X., Li, Q., Ren, C., Guo, W., Li, C., He, Q., Liu, R., & Zou, J. (2019). Real-value negative selection over-sampling for imbalanced data set learning. *Expert Systems with Applications*, 129, 118–134. <https://doi.org/10.1016/j.eswa.2019.04.011>
- Thanathamath, P., & Lursinsap, C. (2013). Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques. *Pattern Recognition Letters*, 34(12), 1339–1347. <https://doi.org/10.1016/j.patrec.2013.04.019>
- Theissler, A., Pérez-Velázquez, J., Kettelgerdes, M., & Elger, G. (2021). Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. *Reliability Engineering & System Safety*, 215, Article 107864.
- Throckmorton, C. S., Mayew, W. J., Venkatachalam, M., & Collins, L. M. (2015). Financial fraud detection using vocal, linguistic and financial cues. *Decision Support Systems*, 74, 78–87.
- Tomek, I. (1976). Two modifications of cnn. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-6(11), 769–772. <https://doi.org/10.1109/TSMC.1976.4309452>
- Tong, L. I., Chang, Y. C., & Lin, S. H. (2011). Determining the optimal re-sampling strategy for a classification model with imbalanced data using design of experiments and response surface methodologies. *Expert Systems with Applications*, 38(4), 4222–4227. <https://doi.org/10.1016/j.eswa.2010.09.087>
- Torgo, L., & Torgo, M. L. (2013). Package 'dmwr'. *comprehensive R archive. Network*.
- Triguero, I., Galar, M., Vluymans, S., Cornelis, C., Bustince, H., Herrera, F., & Saeys, Y. (2015). Evolutionary undersampling for imbalanced big data classification. In *2015 IEEE Congress on Evolutionary Computation*. <https://doi.org/10.1109/CEC.2015.7256961>
- Triguero, I., González, S., Moyano, J. M., García, S., Alcalá-Fdez, J., Luengo, J., Fernández, A., del Jesús, M. J., Sánchez, L., & Herrera, F. (2017). KEEL 3.0: An open source Software for multi-stage analysis in data mining. *International Journal of Computational Intelligence Systems*, 10(1), 1238. <https://doi.org/10.2991/ijcis.10.1.82>
- Tsai, C.-F.-F., Lin, W.-C.-C., Hu, Y.-H.-H., & Yao, G.-T.-T. (2019). Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. *Information Sciences*, 477, 47–54. <https://doi.org/10.1016/j.ins.2018.10.029>
- Vairetti, C., Assadi, J. L., & Maldonado, S. (2024). Efficient hybrid oversampling and intelligent undersampling for imbalanced big data classification. *Expert Systems with Applications*, 246, Article 123149. <https://doi.org/10.1016/j.eswa.2024.123149>
- Verbiest, N., Ramentol, E., Cornelis, C., & Herrera, F. (2014). Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection. *Applied Soft Computing Journal*, 22, 511–517. <https://doi.org/10.1016/j.asoc.2014.05.023>
- Visa, S., & Ralescu, A. (2005). Issues in mining imbalanced data sets-a review paper. *Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference*.
- Vluymans, S., Triguero, I., Cornelis, C., & Saeys, Y. (2016). EPRENNID: An evolutionary prototype reduction based ensemble for nearest neighbor classification of imbalanced data. *Neurocomputing*, 216, 596–610. <https://doi.org/10.1016/j.neucom.2016.08.026>
- Wang, Q. (2014). A hybrid sampling SVM approach to imbalanced data classification. *Abstract and Applied Analysis*, 2014, 1–7. <https://doi.org/10.1155/2014/972786>
- Wang, S., & Yao, X. (2009). Diversity analysis on imbalanced data sets by using ensemble models. In *2009 IEEE Symposium on Computational Intelligence and Data Mining*. <https://doi.org/10.1109/CIDM.2009.4938667>
- Wang, X., Chen, Y., Toth, G., Manchester, W. B., Gombosi, T. I., Hero, A. O., Jiao, Z., Sun, H., Jin, M., & Liu, Y. (2020). Predicting solar flares with machine learning: Investigating solar cycle dependence. *The Astrophysical Journal*, 895(1), 3.
- Wasikowski, M., & Chen, X. W. (2010). Combating the small sample class imbalance problem using feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1388–1400. <https://doi.org/10.1109/TKDE.2009.187>
- Wei, G., Mu, W., Song, Y., & Dou, J. (2022). An improved and random synthetic minority oversampling technique for imbalanced data. *Knowledge-Based Systems*, 248, Article 108839. <https://doi.org/10.1016/j.knsys.2022.108839>
- Weiss, G. M. (2004). Mining with rarity. *ACM SIGKDD Explorations Newsletter*, 6(1), 7–19. <https://doi.org/10.1145/1007730.1007734>
- Weiss, G. M. (2010). The impact of small disjuncts on classifier Learning. *Data Mining*, 8, 193–226. [https://doi.org/10.1007/978-1-4419-1280-0\\_9](https://doi.org/10.1007/978-1-4419-1280-0_9)
- Weng, C. G., & Poon, J. (2006). A data complexity analysis on imbalanced datasets and an alternative imbalance recovering strategy. *Proceedings - 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)*, WI'06, 270–276. <https://doi.org/10.1109/WI.2006.9>
- Wilk, S., Stefanowski, J., Wojciechowski, S., Farion, K. J., & Michalowski, W. (2016). Application of preprocessing methods to imbalanced clinical data: An experimental study. In *Advances in Intelligent Systems and Computing* (Vol. 471, pp. 503–515). [https://doi.org/10.1007/978-3-319-39796-2\\_41](https://doi.org/10.1007/978-3-319-39796-2_41)
- Wilson, D. L. (1972). Asymptotic properties of Nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man and Cybernetics*, 2(3), 408–421. <https://doi.org/10.1109/TSMC.1972.4309137>
- Wong, G. Y., Leung, F. H. F., & Ling, S. H. (2013). A novel evolutionary preprocessing method based on over-sampling and under-sampling for imbalanced datasets. *IECON Proceedings (Industrial Electronics Conference)*, 2354–2359. <https://doi.org/10.1109/IECON.2013.6699499>
- Wong, G. Y. Y., Leung, F. H. F. H., & Ling, S.-H.-S.-H.-H. (2018). A hybrid evolutionary preprocessing method for imbalanced datasets. *Information Sciences*, 454–455, 161–177. <https://doi.org/10.1016/j.ins.2018.04.068>
- Xiong, H., Wu, J., & Liu, L. (2010). Classification with ClassOverlapping: A systematic study. In *Proceedings of the 2010 International Conference on E-Business Intelligence*. <https://doi.org/10.2991/icebi.2010.43>
- Yao, J., Zhang, J., & Wang, L. (2018). A financial statement fraud detection model based on hybrid data mining methods. In *2018 International Conference on Artificial Intelligence and Big Data*. <https://doi.org/10.1109/ICAIBD.2018.8396167>
- Yao, Z., Lum, Y., Johnston, A., Mejia-Mendoza, L. M., Zhou, X., Wen, Y., Aspuru-Guzik, A., Sargent, E. H., & Seh, Z. W. (2023). Machine learning for a sustainable energy future. *Nature Reviews Materials*, 8(3), 202–215.
- Yen, S.-J.-S.-J., & Lee, Y.-S.-S.-Y.-S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36, 5718–5727. <https://doi.org/10.1016/j.eswa.2008.06.108>
- Ying, X. (2019). An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168(2). <https://doi.org/10.1088/1742-6596/1168/2/022022>
- Yoon, K., & Kwek, S. (2005). An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in functional genomics. *Proceedings - HIS 2005: Fifth International Conference on Hybrid Intelligent Systems*, 2005, 303–308. <https://doi.org/10.1109/ICHIS.2005.23>
- Yu, H., Ni, J., & Zhao, J. (2013). ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data. *Neurocomputing*, 101, 309–318. <https://doi.org/10.1016/j.neucom.2012.08.018>
- Yu, S., Guo, J., Zhang, R., Fan, Y., Wang, Z., & Cheng, X. (2022). A re-balancing strategy for class-imbalanced classification based on instance difficulty. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. <https://doi.org/10.1109/CVPR52688.2022.00017>
- Zhai, J., Qi, J., & Shen, C. (2022). Binary imbalanced data classification based on diversity oversampling by generative models. *Information Sciences*, 585, 313–343. <https://doi.org/10.1016/j.ins.2021.11.058>
- Zhang, C., Bi, J., Xu, S., Ramentol, E., Fan, G., Qiao, B., & Fujita, H. (2019). Multi-imbalance: An open-source software for multi-class imbalance learning. *Knowledge-Based Systems*, 174, 137–143. <https://doi.org/10.1016/j.knsys.2019.03.001>
- Zhang, C., Tan, K. C., Li, H., & Hong, G. S. (2019). A cost-sensitive deep belief network for imbalanced classification. *IEEE Transactions on Neural Networks and Learning Systems*, 30(1), 109–122. <https://doi.org/10.1109/TNNLS.2018.2832648>
- Zhang, H. (2017). mixup: Beyond empirical risk minimization. *ArXiv Preprint. ArXiv: 1710.09412*.
- Zhang, H., & Li, M. (2014). RWO-sampling: A random walk over-sampling approach to imbalanced data classification. *Information Fusion*, 20(1), 99–116. <https://doi.org/10.1016/j.inffus.2013.12.003>
- Zhang, Y., Shi, X., Zhang, H., Cao, Y., & Terzija, V. (2022). Review on deep learning applications in frequency analysis and control of modern power system. *International Journal of Electrical Power & Energy Systems*, 136, Article 107744.
- Zhu, H., Zhou, M. C., Liu, G., Xie, Y., Liu, S., & Guo, C. (2024). NUS: Noisy-sample-removed undersampling scheme for imbalanced classification and application to credit Card fraud detection. *IEEE Transactions on Computational Social Systems*, 11(2), 1793–1804. <https://doi.org/10.1109/TCSS.2023.3243925>
- Zhu, R., Hu, X., Hou, J., & Li, X. (2021). Application of machine learning techniques for predicting the consequences of construction accidents in China. *Process Safety and Environmental Protection*, 145, 293–302. <https://doi.org/10.1016/j.psep.2020.08.006>