



## Review

# Overview of machine learning in class imbalance scenarios: Trends, challenges, and approaches

Gilberto Sussumu Hida <sup>a,b,\*</sup>, André Câmara Alves Do Nascimento <sup>a,c</sup>

<sup>a</sup> Cesar School, Recife, 04575-020, Pernambuco, Brazil

<sup>b</sup> Albert Einstein Israelite Hospital - E.GATE, São Paulo, 05652-900, São Paulo, Brazil

<sup>c</sup> Cesar School, UFRPE - Federal Rural University of Pernambuco, Recife, 52171-900, Pernambuco, Brazil

## ARTICLE INFO

## Keywords:

Class imbalance  
Data-level techniques  
Algorithm-level techniques  
Hybrid techniques  
Systematic mapping  
Systematic review of literature

## ABSTRACT

This study presents a systematic mapping of machine learning in class imbalance scenarios, offering a broad overview of key challenges, promising emerging techniques, and established methodologies across various application domains. The investigation stands out by employing a hybrid search and selection protocol that combines methodological rigor with technical innovation.

The adopted strategy integrated manual searches in reference sources with automated processes based on machine learning, semantic embeddings, and graph-based ranking algorithms. To enhance selection quality, the Quasi-Golden Set (QGS) method was used to build a reference set from manually selected articles – a critical foundation for calibrating and evaluating automated search strings. This combination ensured broad coverage of the topic while improving sensitivity and precision in identifying relevant studies.

The initial analysis reviewed 25,593 publications. After screening and applying eligibility criteria, 468 articles were included in the final dataset. The results indicate that 55 % of the studies address multiple domains, with a strong predominance of tabular data (84 %). SMOTE and hybrid approaches were among the most common techniques, present in 61 % of the studies. In terms of evaluation metrics, ROC-AUC was the most frequently used, followed by F1-score and accuracy – the latter noted for limitations in highly imbalanced scenarios. Building on these findings, we derive an empirically grounded taxonomy that links problem context, solution algorithms, and scenario-appropriate evaluation metrics, and we provide a minimal selection guideline table to support applied use.

While sampling-based methods remain prevalent, deep learning approaches such as convolutional neural networks and graph-based models are increasingly adopted. Additionally, federated, contrastive, and semi-supervised learning are emerging as relevant paradigms, particularly suited for privacy-aware or low-label environments.

This study consolidates current knowledge, identifies methodological and application gaps, and highlights trends that are likely to shape future research. It contributes both a comprehensive synthesis of the field and strategic insights for advancing machine learning techniques in the presence of class imbalance.

## 1. Introduction

Class imbalance is a significant challenge in various data science domains, such as computer vision, natural language processing, and tabular data analysis, Ghosh et al. (2024), Johnson and Khoshgoftaar (2019), Lipitakis and Lipitakis (2014). This issue arises when one or more classes are underrepresented compared to others, causing machine learning models to favor the majority classes. As a result, the accuracy and generalizability of the models can be compromised (Johnson & Khoshgoftaar, 2019), particularly in critical applications where accurate predictions

for minority classes are crucial, such as fraud detection, medical diagnosis (Ragonesi et al., 2023), and security analysis.

Imbalanced learning is a widely studied topic in the scientific literature, evidenced by a substantial volume of publications and the organization of workshops at renowned conferences such as AAAI (Association for the Advancement of Artificial Intelligence) and ICML (International Conference on Machine Learning) (Chawla et al., 2002; Provost, 2000).

Several systematic reviews and mapping studies have been conducted to synthesize the accumulated knowledge in this field (Buda et al., 2018; Felix & Lee, 2019; Johnson & Khoshgoftaar, 2019;

\* Corresponding author.

E-mail addresses: [gsh@cesar.school](mailto:gsh@cesar.school), [gilberto.hida@einstein.br](mailto:gilberto.hida@einstein.br), [sussumu.educacional@gmail.com](mailto:sussumu.educacional@gmail.com) (G.S. Hida), [acm@cesar.school](mailto:acm@cesar.school), [andre.camara@ufrpe.br](mailto:andre.camara@ufrpe.br) (A.C. Alves Do Nascimento).

<https://doi.org/10.1016/j.eswa.2025.129592>

Received 28 November 2024; Received in revised form 8 August 2025; Accepted 30 August 2025

Available online 2 September 2025

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

Kaur et al., 2019; Shakeel et al., 2017; Sneha & Annappa, 2024; Susan & Kumar, 2021; Werner de Vargas et al., 2023; Wang et al., 2021). This study differentiates itself from previous works by covering a significant number of articles and applying rigorous systematic criteria in the search and selection stages, thus minimizing the biases inherent in secondary studies. This robust methodological approach ensures a more comprehensive and impartial analysis of the state-of-the-art in imbalanced learning.

This systematic mapping aims to provide a comprehensive overview of the challenges, methodologies, and potential knowledge gaps in the area of class imbalance. The approaches for training models in this context can be grouped into three main categories: data-level, algorithm-level, and hybrid approaches (de Moraes et al., 2016; Li et al., 2020; Wu & Li, 2024). At the data level, techniques such as oversampling, undersampling, and the use of Generative Adversarial Networks (GANs) are commonly used to balance classes before model training (Lipitakis & Lipitakis, 2014). At the algorithm level, strategies such as adjusting class weights, modifying loss functions, and developing cost-sensitive algorithms are frequently used (Ochal et al., 2023; Zhou et al., 2024). Hybrid approaches integrate techniques from both levels to leverage the unique advantages of each, aiming to maximize their combined benefits (Shi et al., 2023; Wu & Li, 2024).

To guide the reader through the remainder of this work, the structure of the article is organized as follows: Section 2 reviews related works, highlighting prior surveys and mappings on class imbalance. Section 3 presents the main conceptual foundations and challenges of imbalanced learning. Section 4 details the methodology adopted, including the hybrid search strategy and selection procedures. Section 5 outlines the results obtained from the mapping process, while Section 6 offers a bibliometric analysis of the selected literature. Section 7 discusses the findings in light of the four research questions. Section 8 introduces an empirically derived taxonomy (problem context → Algorithms → evaluation) together with a minimal, scenario-based guideline (user cases) to support method selection and reporting. Section 9 concludes the study, summarizing key insights and directions for future research.

## 2. Related works

This section provides an overview of existing secondary studies on class imbalance in machine learning, including both literature reviews and systematic mappings. The objective is to position the present work within the context of prior syntheses and to highlight its distinctive contributions.

### 2.1. Reviews and surveys

Susan and Kumar (Susan & Kumar, 2021) (2021) and (Shakeel et al., 2017) (2017) conducted surveys that explored the challenges and commonly used techniques for handling imbalanced data. In Susan and Kumar (2021), the authors focused on intelligent sampling techniques and the use of evolutionary algorithms, while in Shakeel et al. (2017) provided a detailed analysis of methods based on neural networks and SVM.

Several reviews and surveys have been conducted that focus on more specific methods, such as Buda et al. (2018) (2018), which investigates the impact of class imbalance on classification performance in Convolutional Neural Networks (CNNs). Furthermore, Johnson and Khoshgof-taar (2019) (2019) conducted a systematic review, analyzing 15 studies that examine existing deep learning techniques to address mismatched data.

Felix and Lee (2019) (2019) conducted a systematic review following Kitchenham's guidelines (Kitchenham et al., 2007). The analysis included 118 articles, selected from an initial pool of 1,673. One of the main findings was the need for adequate data pre-processing. The study also emphasized the scarcity of formal systematic reviews in the field,

noting that only 2% of the existing reviews adhered to rigorous systematic review methodologies.

Kaur et al. (2019) (2019) presented a review of methods and challenges in the field of imbalanced data. Of 385 articles initially assessed, 152 were selected for analysis. The study noted the prevalence of data modification techniques and highlighted the potential of deep learning as a promising area for future research.

Wang et al. (2021) (2021) provided a detailed examination of the main approaches to classifying imbalanced data, highlighting sampling methods, algorithmic techniques, cost-sensitive methods, and deep learning approaches.

Werner de Vargas et al. (2023) (2023) conducted a systematic mapping focused on data preprocessing techniques, analyzing 9927 articles from seven digital libraries, of which 35 were selected for in-depth analysis. The study mainly focused on sampling techniques, with an emphasis on the predominance of oversampling. As future research directions, the study suggests the development of hybrid approaches.

Sneha and Annappa (2024) (2024) reviewed key approaches in the field, with the added contribution of discussing evaluation metrics, offering a more critical analysis of model effectiveness.

Hairani et al. (2024) (2024) conducted a systematic review focused on SMOTE and its variations applied to the health context. The study analyzed 70 articles published between 2019 and 2023, aiming to investigate and evaluate different modifications proposed for the SMOTE method to address data imbalance. The review highlighted that approaches combining SMOTE with noise filtering techniques, clustering strategies to minimize class overlap, and adjustments based on distance metrics can enhance the effectiveness of handling imbalanced data, thereby improving the accuracy and robustness of models in healthcare scenarios.

The proposed study differentiates itself from previous works through its extensive coverage of mapped articles, combined with the application of a rigorous search and selection methodology.

The timeline in Fig. 1 summarizes the publication dates of reviews and surveys in the field.

### 2.2. Contribution

As highlighted in previous reviews and surveys (Felix & Lee, 2019; Hairani et al., 2024; Werner de Vargas et al., 2023), the field of imbalanced learning has shown significant growth in the volume of publications. However, there is still a scarcity of studies that systematically organize and synthesize these primary works. When examined from the perspective of methodologies proposed for systematic reviews in computing, the number of studies adhering to such guidelines is even smaller.

The systematic mapping proposed in this study stands out by exploring methodologies applied across multiple domains, offering a comprehensive and comparative view of the field by analyzing not only methodologies and applications but also the growth of the research area focusing on the main trends. The study adopts a rigorous approach aligned with formal methodologies recommended for systematic reviews in computing (Kitchenham et al., 2007; Petersen et al., 2015) and introduces new tools in the search and selection stages to mitigate the inherent subjectivity of these processes. The search coverage is extensive, mapping a total of 25593 articles (compared to 9927 articles reviewed in Werner de Vargas et al., 2023), resulting in a final selection of 468 studies (in contrast with 152 articles selected in Kaur et al., 2019).

Beyond consolidating trends and methods, this study contributes: (i) an empirically derived taxonomy that integrates task type, data modality, imbalance characteristics, learning regime, solution algorithms (data-level, algorithm-level, hybrid), and evaluation categories (Section 8.1); and (ii) a minimal, scenario-based selection guideline that operationalizes what, when and how to use and evaluate (Section 8.2, Table 4).

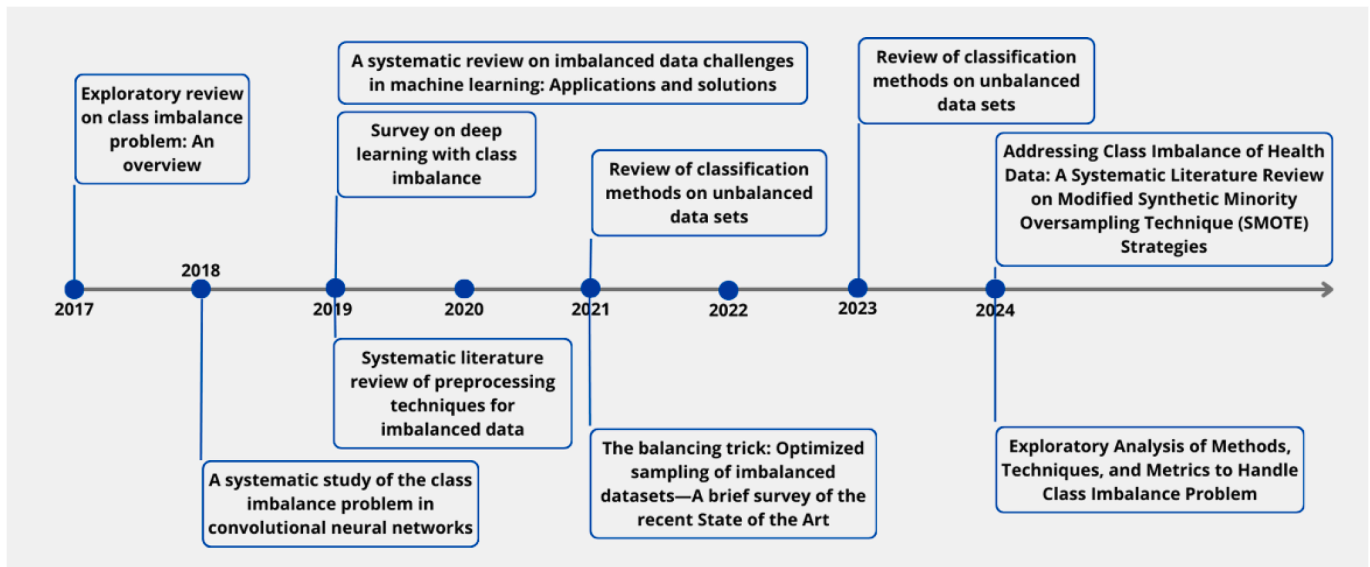


Fig. 1. Timeline of surveys and reviews.

### 3. Preliminaries

Machine learning in the presence of imbalanced data imposes additional challenges due to the skewed nature of the class distribution, which can lead to models favoring the majority classes (Ochal et al., 2023), (Li et al., 2020), and (He & Garcia, 2009). This occurs because machine learning algorithms, especially supervised ones, tend to optimize their cost functions based on metrics like overall accuracy, which is heavily influenced by the majority class. As a result, models can overlook or misclassify minority classes, which are often the most critical in sensitive contexts. Some of the additional challenges mentioned in the literature are as follows.

- **Algorithm Functionality:** Standard learners (e.g., decision trees, neural networks, SVMs) often underperform on imbalanced data: optimizing overall accuracy biases them toward the majority and harms minority generalization (He & Garcia, 2009).
- **Model Instability:** Imbalanced data can destabilize training—especially in deep networks—because majority-dominated gradients drive suboptimal convergence that overlooks minority structure (Buda et al., 2018; He & Garcia, 2009; Werner de Vargas et al., 2023; Wang et al., 2021).
- **Inadequate Evaluation Metrics:** Under class imbalance, overall accuracy is an unreliable indicator of performance; reliance on it can produce misleading conclusions about model quality (Chawla et al., 2002; He & Garcia, 2009; Saito & Rehmsmeier, 2015).

Methods for handling class-imbalance effects are commonly grouped into three categories—data-level, algorithm-level, and hybrid (Sneha & Annappa, 2024)—and, rather than enumerating variants here, we use this triad merely as background and present in Section 8.1 a new empirically derived taxonomy that guides method selection by problem context and evaluation.

### 4. Methods and procedures

In this section, we presented the structure adopted for the development of the study, from the formulation of the research questions to the final stage of data extraction from the selected studies.

This section is divided into eight key topics.

1. **Research Questions:** We begin by addressing the questions that this study seeks to answer.

2. **Overview:** This part provides a general description of the search and selection process flow.
3. **Search Strategy:** Here, we detail the systematic method adopted for the search phase.
4. **Databases:** A description of the sources consulted.
5. **Inclusion and exclusion criteria:** Presents the inclusion and exclusion criteria of the articles at all stages of the systematic mapping protocol.
6. **Selection Methods:** The structure for the final selection of articles that will compose the study set is presented in detail.
7. **Data Collection:** The results obtained from applying the search and selection processes.
8. **Limitations:** A discussion of the study's limitations and potential biases of the study.

#### 4.1. Research questions

The specific research questions of this systematic mapping are the following:

- RQ1 What are the main challenges faced when dealing with imbalanced data in machine learning models?
- RQ2 What techniques and methodologies have been most commonly used to address class imbalance, and what are their advantages and disadvantages?
- RQ3 In which application domains are these techniques most commonly applied?
- RQ4 What are the emerging trends and future research directions in the field of class imbalance in machine learning?

Among the research questions, RQ4 stands out as the central focus of this study, given its emphasis on emerging trends and future directions. Accordingly, it received the most comprehensive treatment throughout the manuscript.

#### 4.2. General description of the search and selection process

This study is characterized as a systematic mapping, a rigorous methodological approach that aims to provide a comprehensive overview of a given research field, identifying gaps and trends in the existing literature. According to the definitions proposed by Kitchenham and Charters (Kitchenham et al., 2007), a systematic mapping organizes

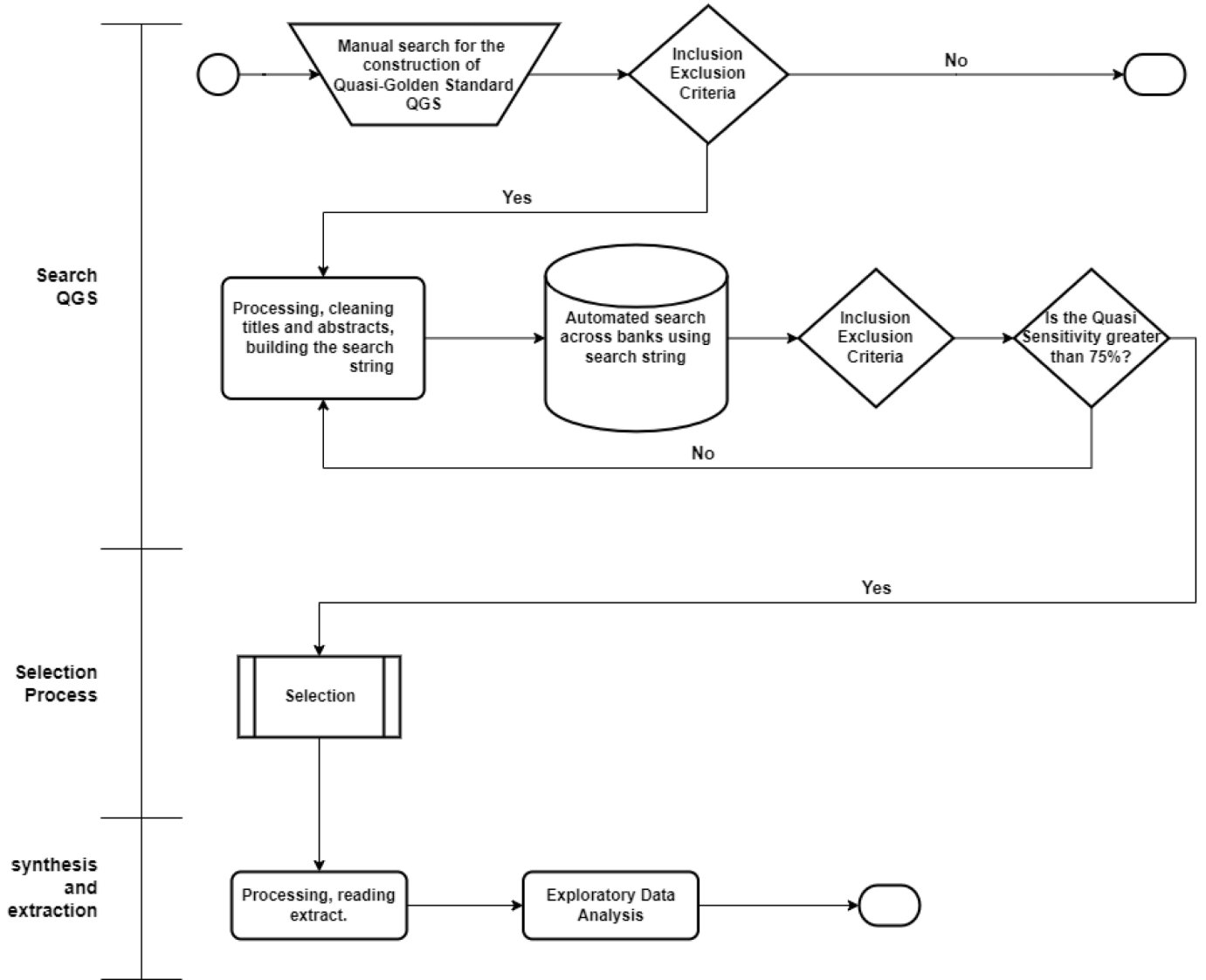


Fig. 2. General structure of the study's search, selection, and extraction process.

and structures the literature of a specific area of study, categorizing studies by topics, methods used, and application areas, among other criteria (Kitchenham et al., 2007; Petersen et al., 2008). The general structure of the study is represented in the flow chart (Fig. 2).

#### 4.3. Search strategy

The search strategy employed in this study is a hybrid approach known as QGS (Quasi-Golden Standard), which combines manual and automated search procedures (Zhang & Ali Babar, 2010; Zhang et al., 2011). The diagram in Fig. 3 illustrates the QGS methodology in detail, corresponding to the initial phase of the overall study framework depicted in Fig. 2.

As shown in Fig. 3, the QGS method involved creating a reference set of known studies through manual searches in key sources, which was then used to develop and evaluate search strings in automated searches. These strings were iteratively refined until a satisfactory pseudo-sensitivity (Zhang & Ali Babar, 2010) was achieved, ensuring that the search efficiently retrieved the maximum number of relevant studies.

The sources for the manual search included the six most influential journals in Machine Learning, which were selected based on the 2023 Journal Citation Reports (JCR), along with the four most relevant conferences in the field, which were identified by the Google Scholar h5 index.

To address RQ3, we considered all available articles from the selected journals and conferences. After constructing the Quasi-Golden Set, we applied Natural Language Processing methods, with an emphasis on TF-IDF (Saeidmehr et al., 2024), to identify the most representative words in the field. This enabled a more precise and effective development of search strings for automated search, optimizing the retrieval of relevant studies (Zhang & Ali Babar, 2010).

The pseudo-sensitivity threshold was established between 75 % and 85 % as indicated in previous studies (Zhang & Ali Babar, 2010; Zhang et al., 2011). The major advantage of the Quasi-Golden Set (QGS) search method lay in the evaluation of search quality, as well as in establishing a systematic approach to the data collection phase. Throughout the text, the acronym QGS is used to denote both the methodology and the set of articles retrieved through it; we believe the context in which the acronym appears makes its intended meaning clear in each instance.



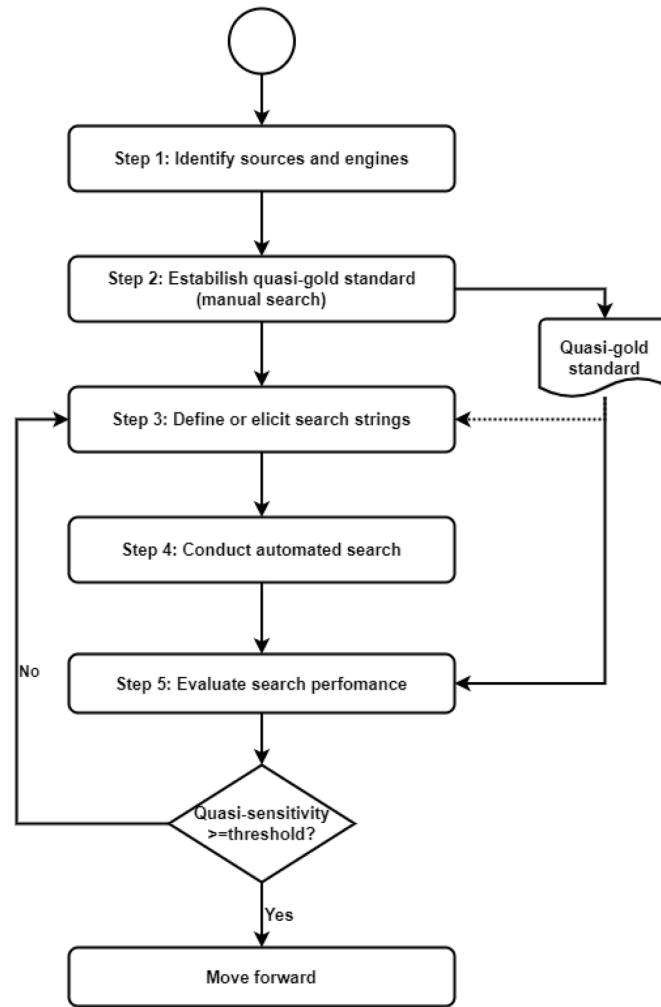


Fig. 3. Flowchart of the QGS search method (Zhang & Ali Babar, 2010).

#### 4.4. Databases

For the manual search, we used the following journals:

1. IEEE Transactions on Neural Networks and Learning Systems.
2. Journal of Machine Learning Research (JMLR).
3. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).
4. Machine Learning Journal (Springer).
5. Pattern Recognition.
6. Neural Networks (Elsevier).

The following conferences will be included:

1. Conference on Neural Information Processing Systems (NeurIPS).
2. International Conference on Machine Learning (ICML).
3. Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI).
4. International Conference on Artificial Intelligence and Statistics (AISTATS).

For the automated search phase, we used the databases:

1. IEEE Xplorer.
2. ACM Digital Library,
3. Scopus.
4. Web of Science.

#### 4.5. Inclusion and exclusion criteria

- Inclusion: The selected articles met the following set of attributes:
  - Primary studies.
  - Studies in the English language.
  - Studies published in the following types of publications: journals, transactions, conference papers, workshop papers, and proceedings papers.
- Exclusion: Articles were excluded if they exhibited any of the following characteristics:
  - Book chapters, short papers, or case reports.
  - Studies focusing on applications rather than methods for handling class imbalance.

Both inclusion and exclusion criteria were applied during the construction of the QGS as well as in the automated search phase.

#### 4.6. Selection methods

We will present the structure of the selection process, which involves the use of machine learning methods combined with manual inspections to ensure both the reliability and effectiveness of the approach, while also allowing for its scalability.

After the execution of the automated search in the aforementioned databases, the next step was to select the articles relevant to the theme, with the aim of answering the defined research questions. Given the large volume of articles recovered, a hybrid selection methodology was

adopted, relying heavily on machine learning algorithms to assist in the selection of the studies (Ferreira et al., 2021; Octaviano et al., 2015; Popoff et al., 2020; Roth & Wermer-Colan, 2023; Van De Schoot et al., 2021).

Initially, duplicate removal, language restriction, and selection of research and conference papers were performed. To improve the precision of the selection process, semantic embeddings (Mikolov, 2013) were used to capture the relationships between words and phrases, facilitating comparison between the titles and abstracts of articles. This enabled the creation of a vector representation of the abstracts, a critical step for the application of machine learning algorithms. The chosen model for generating these vectors was OpenAI's "text-embedding-3-small", which produces representations in a 1536-dimensional vector space.

In the subsequent step, three classifiers were constructed to determine whether an article should be included in the final selection based on its relevance to the proposed research questions.

- **First Decider:** A machine learning algorithm, Random Forest, will be trained.

The training set was constructed using the QGS and a sample from the dataset obtained through the automated search, following manual inspection.

The model evaluation methods included cross-validation, with the ROC-AUC metric being the primary evaluation measure.

- **Second Decider:** A similarity matrix with the articles that make up the QGS.

Given the adoption of embeddings to capture the semantic relationships between articles, cosine similarity was chosen as the main metric to measure the proximity between text vectors. This metric was widely used due to its ability to compare texts, even when they employed different terminologies, capturing semantic similarity in a robust manner. The application of cosine similarity allowed the identification of articles with related content, even if their linguistic expressions varied significantly.

- **Third Decider:** The third decider was based on the use of a Large Language Model (LLM) to determine whether an article should be included in the study. The adopted model was GPT-3.5-turbo.

The adoption of LLMs at this stage aims to ensure a more sophisticated evaluation of the articles, leveraging the ability of these models to interpret linguistic and semantic nuances comprehensively, thereby increasing the precision in selecting the most relevant studies. Some studies (Petersen & Gerken, 2024; Scherbakov et al., 2024) already highlight the use of LLMs in various phases of systematic mapping/review processes.

After the vote of each decider, cases with unanimous decisions—either for exclusion (0 votes) or inclusion (3 votes) – were considered straightforward and did not undergo manual inspection. Cases with vote divergence (1 or 2 votes), due to their complexity and the need for a more detailed evaluation, were submitted to manual inspection for the final decision.

Finally, a graph analysis was conducted to reduce the cardinality of the final set of articles and eliminate possible redundancies (Voudigari et al., 2016). Fig. 4 presents the flow of the selection process.

#### 4.7. Cardinality reduction: Graph analysis

To reduce the volume of articles in the data collection phase, we will employ graph-based methods, aiming to decrease the number of publications analyzed while minimizing the loss of representativeness. We will use embeddings derived from titles and abstracts to obtain the vector representation of the articles. Then, we will calculate the similarity between each article and the others, resulting in a dense graph where each vertex represents an article and each edge denotes the similarity between two articles.

To extract a representative subgraph, we will adopt the TopRank algorithm (Voudigari et al., 2016), a ranking technique specifically

designed for vertices in graphs, derived from PageRank (Page, 1998). Our approach will involve selecting the most representative articles, located at the top of the ranking, and identifying outliers, positioned at the bottom of the ranking. This strategy will allow us to capture both central research trends and emerging frontiers in the field.

#### 4.8. Data collection and extraction

After selecting the articles, the data collection phase will involve a thorough reading and tabulation of the findings. For this stage, we will once again utilize natural language processing (NLP) algorithms to consolidate the identified key points. Specifically, we will adopt TF-IDF (Saeidmehr et al., 2024) and N-Grams (Brown et al., 1988) techniques, which are widely recognized for their effectiveness in text analysis and extraction of relevant information. Subsequently, descriptive statistical methods will be used for the information summarization phase. These approaches will enable a deeper understanding of the trends and patterns present in the selected publications.

#### 4.9. Limitations and biases

The QGS method is highly dependent on the manual search phase (Zhang & Ali Babar, 2010; Zhang et al., 2011). To mitigate the inherent bias in this step, we defined 10 research sources using objective criteria. Another limitation lies in the initial formulation of search strings, which is based on subjective criteria derived from the Quasi-Golden Set and may require additional iterations. To minimize this bias, we will employ Natural Language Processing (NLP) tools in the construction of the search strings.

Another point of consideration is that the method is directed towards maximizing search sensitivity, a critical characteristic for mapping studies (Petersen et al., 2008). However, this can result in the retrieval of numerous irrelevant studies, which will need to be filtered in subsequent analysis stages.

In the selection phase, the use of a machine learning-based selection methodology introduces biases into the study, such as similarity-based selection, which may reduce the diversity of included studies, and the use of an LLM (GPT-3.5-turbo) to which we do not have access to all parameters, adding an element of uncertainty to the process. To counter these effects, manual inspection has been adopted for articles where there was disagreement among the models used.

The use of graphs and ranking methods can introduce representativeness bias, as selection reduces the total number of articles analyzed. To mitigate this bias, we will adopt a sampling strategy that includes both core elements and outliers from the preselected set. This approach will ensure that the diversity of perspectives is captured, allowing for a more comprehensive and representative analysis of the subject under study.

### 5. Search and selection results of the papers

This section presents the results obtained from the search and selection stages described in the previous section. It provides a detailed account of the quantity, distribution, and filtering process of the retrieved articles, including duplicate removal, pseudo-sensitivity analysis, voting outcomes from the selection models, and the application of graph-based cardinality reduction.

#### 5.1. Results of the QGS and search string

The manual search resulted in 244 articles, as shown in the distribution below:

**Duplicate Articles:** A duplicate article was identified, having been published in both IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) in 2021 (18 pages) and in a conference by the Association for the Advancement of Artificial Intelligence (AAAI) in 2017

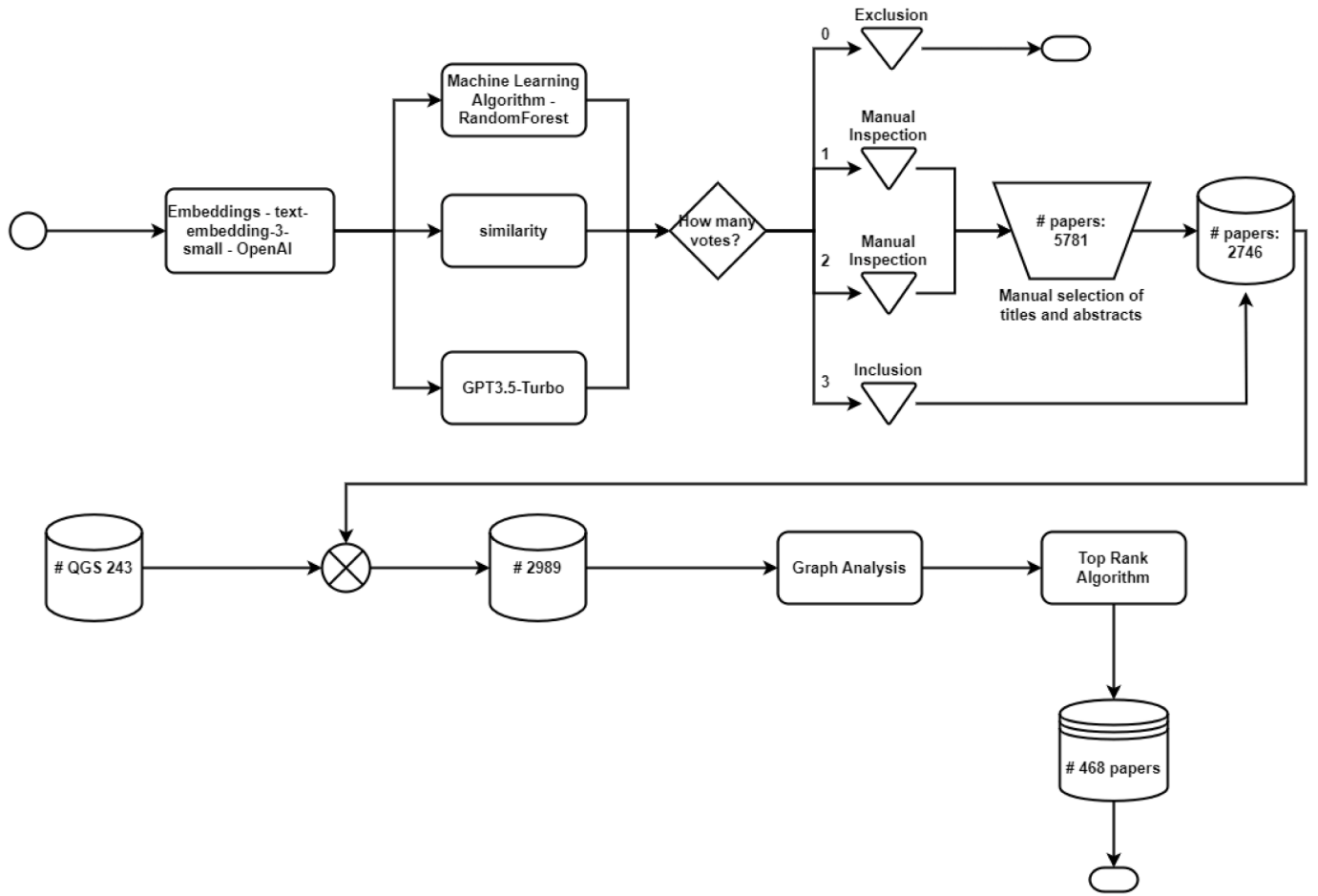


Fig. 4. Flowchart of the selection method.

Table 1  
Results QGS.

Databases	Number of papers
IEEE Transactions on Neural Networks and Learning Systems	36
Journal of Machine Learning Research	3
IEEE Transactions on Pattern Analysis and Machine Intelligence	10
Machine Learning Journal Springer	26
Pattern Recognition	48
Neural Networks	33
NeurIPS	36
ICML	14
AAAI	32
AISTATS	5
Total	243

(7 pages). We chose to retain the most recent version published in IEEE. The final set (QGS) consisted of 243 articles.

Using 2-Grams, we obtained the most important terms, as shown in Fig. 5.

The search string for the automated search was defined as:

```
("imbalanced data" OR "class imbalance" OR
"imbalanced learning" OR "class imbalanced" OR
"imbalanced datasets")
```

## 5.2. Automated search

### 5.2.1. Duplicates

The automated search in the aforementioned databases resulted in a total of 59945 articles, already filtered to include only research and

conference papers. After removing duplicates based on titles, the number of articles was reduced to 25593, representing a decrease of 57.3 % from the initial total. This high reduction rate indicates a considerable overlap among the databases searched.

### 5.2.2. Pseudo-sensitivity

In the QGS method, it is essential to assess the quality of the automated search by verifying how many articles in the QGS were effectively recovered. Of the 243 articles that make up the QGS (Table 1), only 33 were not recovered (primarily articles from Springer), resulting in a total of 210 recovered articles. Thus, the pseudo-sensitivity obtained was 210/243 (0.86). Since this value falls within the established range, it was not necessary to perform another iteration in the search string construction process.

### 5.2.3. Deciders

For the first decider, the training set was constructed by sampling approximately 498 articles from the total dataset, which were inspected and labeled (included or not in the final selection), in addition to the QGS articles, totaling 739 articles (64% exclusion and 36% inclusion).

Cross-validation methods were adopted, with an 80/20 train-test split ratio. The ROC curve of the trained model is presented in Fig. 6.

We obtained a ROC-AUC of 0.97, indicating a good fit.

For the second decider, we used similarity with the QGS database. For each article retrieved by the automated search, we calculate its similarity to the articles in the QGS. The final similarity of an article was defined as the maximum similarity value between it and the QGS articles. Articles with high similarity were selected, with the cutoff point set as the mean plus two standard deviations.

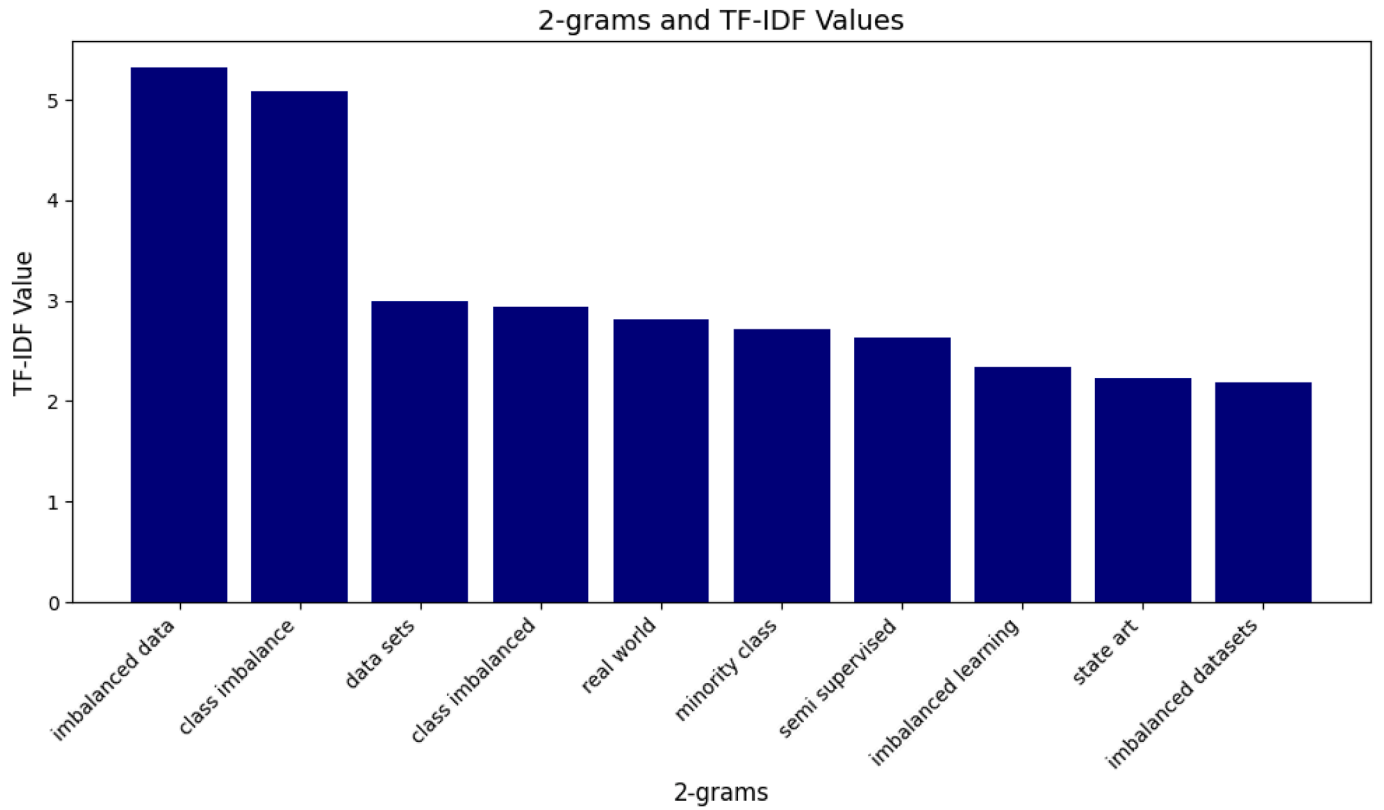


Fig. 5. Most important terms using 2-grams and TF-IDF.

For the third decider, we adopted the following context for analysis.

“You are a contributor to a systematic mapping article on machine learning for imbalanced or unbalanced class learning. The objective is for you to help decide whether an article should be included in the mapping, knowing that we seek articles whose primary focus is learning under these imbalance conditions, rather than those that merely applied a balancing method to solve a particular problem. We aim to select articles that establish challenges, methodologies/approaches to the topic, research directions, and advancements in the field, while ensuring that the core focus of the selected articles remains learning under imbalance conditions. Be judicious, as the articles being analyzed have already passed through filters, so almost all address/use some concept of class imbalance.”

Furthermore, we used 10 evaluation examples to fine-tune the response according to Few-Shot Learning (Perez et al., 2021).

The final results, obtained through voting by the three deciders, are presented in Fig. 7.

A total of 5781 articles received 1 or 2 votes and were therefore manually inspected, with title and abstract analysis to support the decision for inclusion or exclusion. At the end of this stage, 1760 articles were deemed suitable for inclusion. The final composition of selected articles is as follows:

- 1760 articles resulting from manual inspection (articles with 1 or 2 votes).
- 986 articles that received 3 votes for inclusion.
- 243 articles from the QGS set

**Totalling 2989 articles.**

#### 5.2.4. Graph analysis

The use of graph-based ranking methods provided the configuration presented in Fig. 8, where we chose not to represent the edges (as the graph is dense).

Using the TopRank algorithm, we constructed a selection that included the top 236 ranked articles along with the 232 lowest ranked articles, resulting in a total of 468<sup>1</sup> papers selected for the data extraction and collection phase. In Fig. 9, we present the original graph along with the subgraph formed by the articles selected for the next phase of the study.

## 6. Bibliometric analysis

The volume of publications on this topic has shown significant growth, primarily driven by the increasing adoption of machine learning techniques across a wide range of human activities. In particular, 41 % of the articles included in the final set were published from 2022 onward, highlighting an upward trend in the field of imbalanced learning. This growing academic output reflects not only the rising interest in the topic but also the relevance and importance of addressing the challenges associated with data imbalance in machine learning models (Fig. 10).

The growth exhibits a monotonic pattern, indicating stability in this trend; it should be noted that articles selected for 2024 were only available up to August 2024 (Fig. 11).

Regarding publication types, 82 % of the selected articles were published in journals and conferences (Fig. 12).

Finally, there is a notable concentration of publications from institutions in China, which account for 38 % of the selected articles, followed by the United States with 10 %, as shown in Fig. 13.

We observed that certain regional differences persist, among the countries that most contribute to the field. More precisely, when analyzing the topics covered, we found that Japan, South Korea, Taiwan, and Germany exhibit a proportionally higher focus on image-based methods compared to other approaches, as highlighted in Fig. 14.

<sup>1</sup> The full list of analyzed articles is available in the Selected\_Articles\_Set.pdf

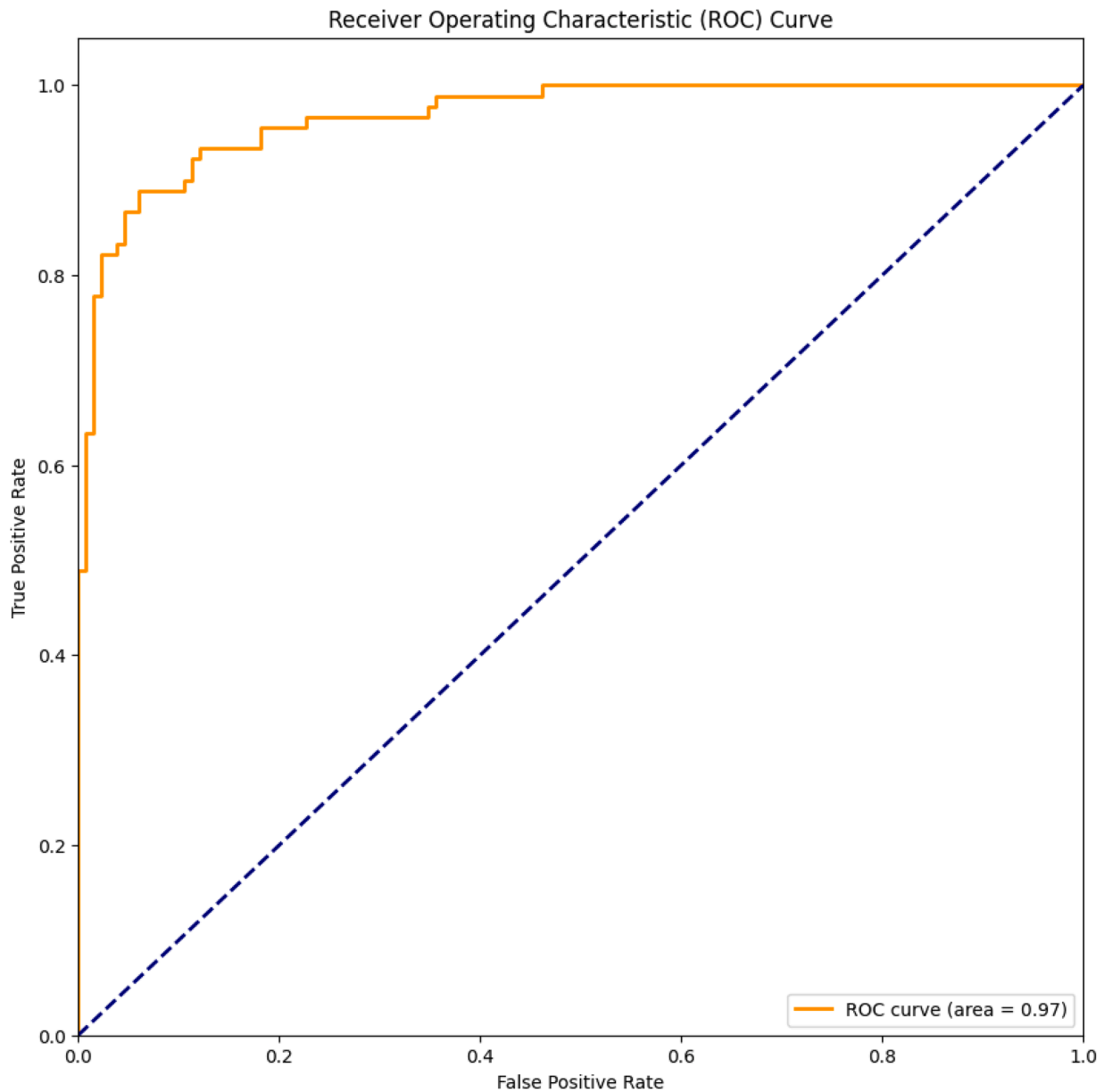


Fig. 6. ROC curve of the trained model.

Applying a more detailed grouping to detect trends and gaps, the following classification was used to categorize each methodology proposed in the papers.

- **Classical Machine Learning:** Approaches based on traditional machine learning models, such as *Naïve Bayes*, *Decision Trees*, *Support Vector Machines (SVM)*, and *Logistic Regression*, among others.
- **Clustering:** Clustering algorithms, including *K-means*, *k-NN*, *DBSCAN*, and other partition-based and density-based approaches.
- **Contrastive Learning:** Techniques focused on maximizing the similarity between examples of the same class while differentiating between different-class examples.
- **Cost-Sensitive:** Methods that incorporate differentiated cost functions and penalty adjustments to mitigate the impact of class imbalance.
- **Deep Learning:** Approaches based on deep neural networks, including architectures such as *GANs*, *CNNs*, *ELMs*, among others.
- **Ensemble Methods:** Techniques that combine multiple models to improve generalization, including *boosting*, *bagging*, *stacking*, and *voting*.

- **Graphs:** Graph-based methods, particularly those leveraging *Graph Neural Networks*.
- **Information Theory:** Application of information theory concepts in learning, such as entropy measures and information-based inference.
- **Sampling Theory:** Sampling-based methods, including oversampling and undersampling techniques.
- **Statistical Methods:** The use of statistical tools for analysis and modeling, such as confidence intervals, *kernel methods*, and *propensity score*.
- **Others:** Miscellaneous methods not covered in the previous categories, including *genetic algorithms*, *mathematical optimization*, and *fuzzy logic*.

Each analyzed study may belong to more than one category. For example, a study that applies *SMOTE* together with deep neural networks to mitigate class imbalance would be classified under both **Sampling Theory** and **Deep Learning**.

As a result, we observed that for the 15 countries with the highest publication output in the field, the distribution of methodological interests is highly heterogeneous, with no uniform pattern (Fig. 15).



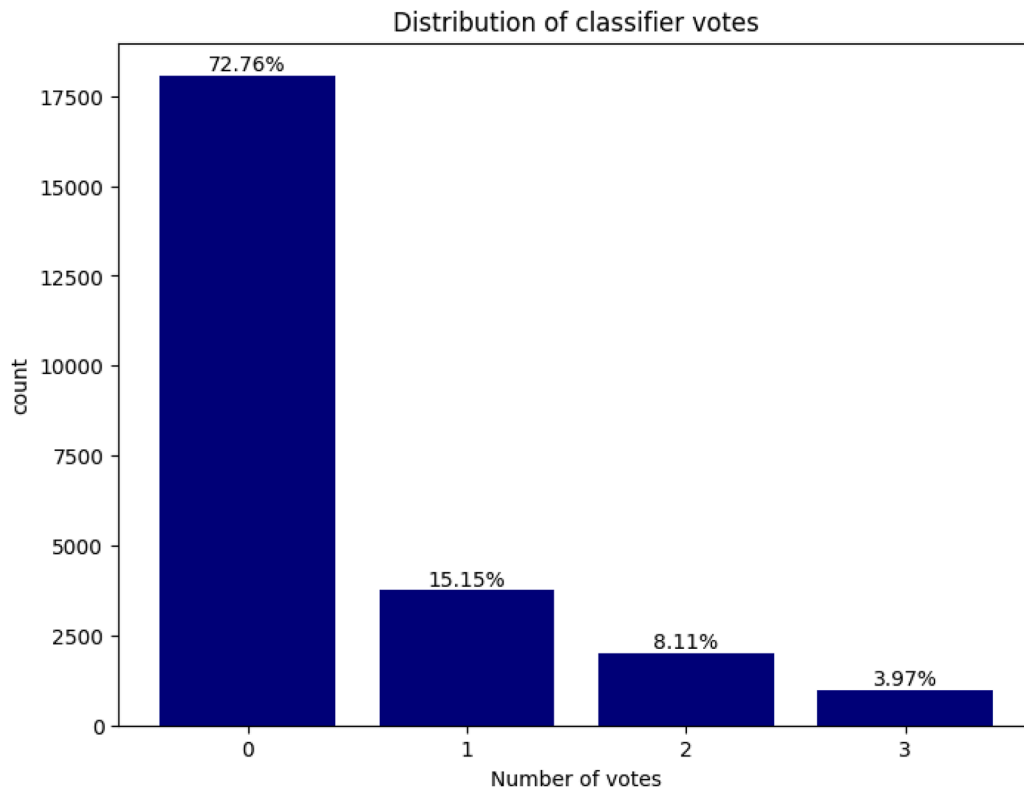


Fig. 7. Voting results by vote count.

## 7. Research questions

After reading and extracting the key terms present in the evidence from the selected articles, we applied NLP methods again, specifically TF-IDF and N-Grams, to conduct a descriptive analysis that supports the answers to the research questions.

*7.1. RQ1: What are the main challenges faced when dealing with imbalanced data in machine learning models?*

The most significant challenges in handling imbalanced data in machine learning models involve a range of complex issues that directly impact algorithm effectiveness. One of the primary problems

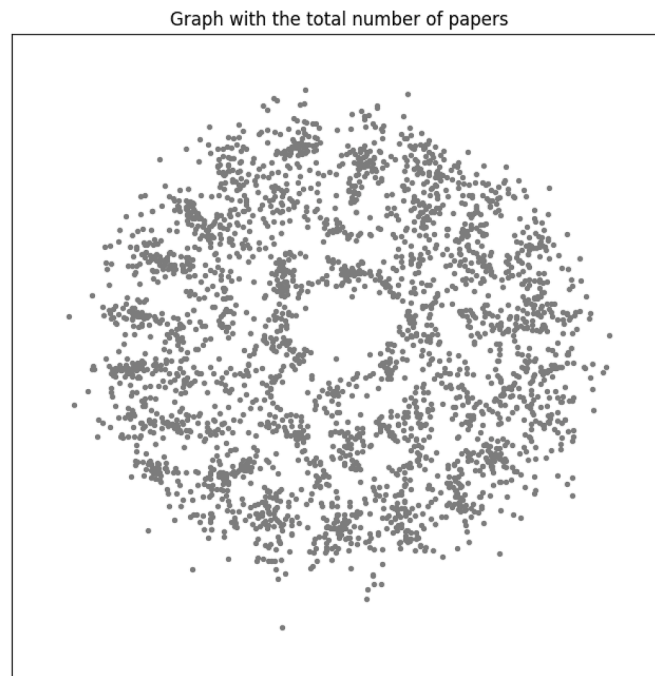


Fig. 8. Representation of the 2989 selected papers.

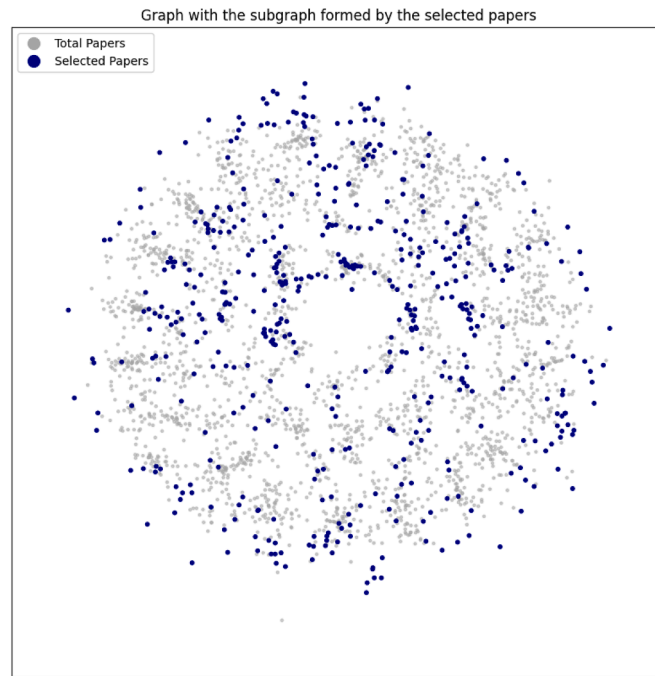
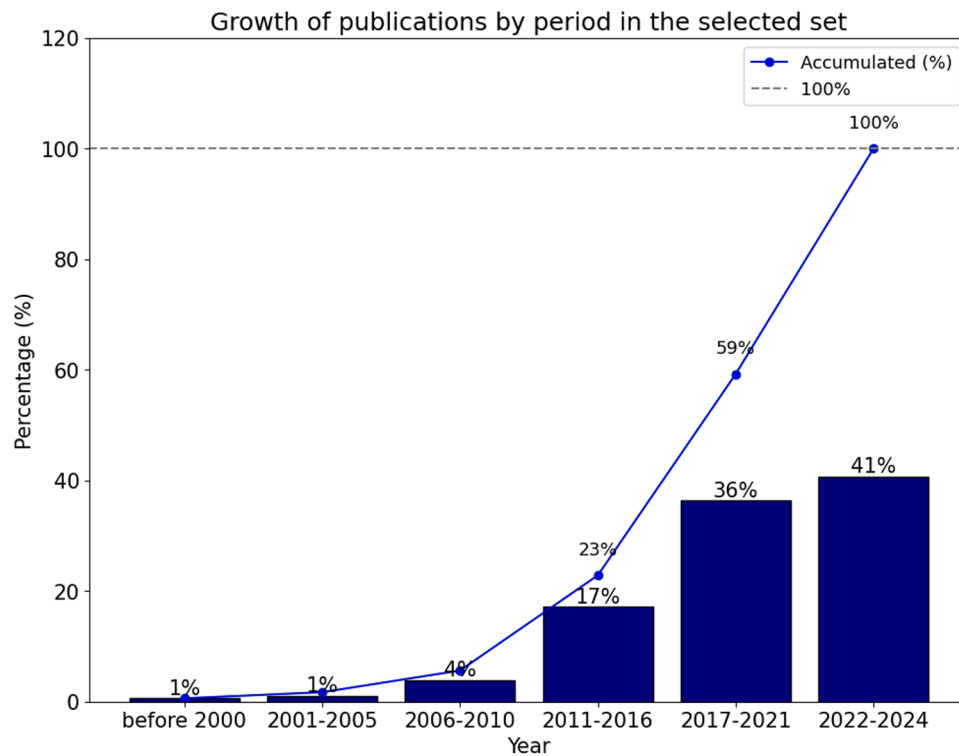


Fig. 9. Representation of the 468 selected articles from a total of 2989 articles.



Obs: For the year 2024, the articles mapped were until August.

Fig. 10. Number of publications over time.

is the bias introduced by the majority class, which can dominate the training process, leading to a low representativeness of the minority class. This issue is often exacerbated by class overlap, which complicates discrimination between classes, and by data distribution inequality, resulting in a skewed configuration that hinders learning.

Furthermore, the presence of diffuse overlapping features can complicate the classification further, making the task of identifying minority samples even more challenging (Fig. 16).

Models frequently encounter difficulties in achieving the desired accuracy, which can lead to overfitting, especially when training datasets are limited in size and diversity. These challenges not only compromise

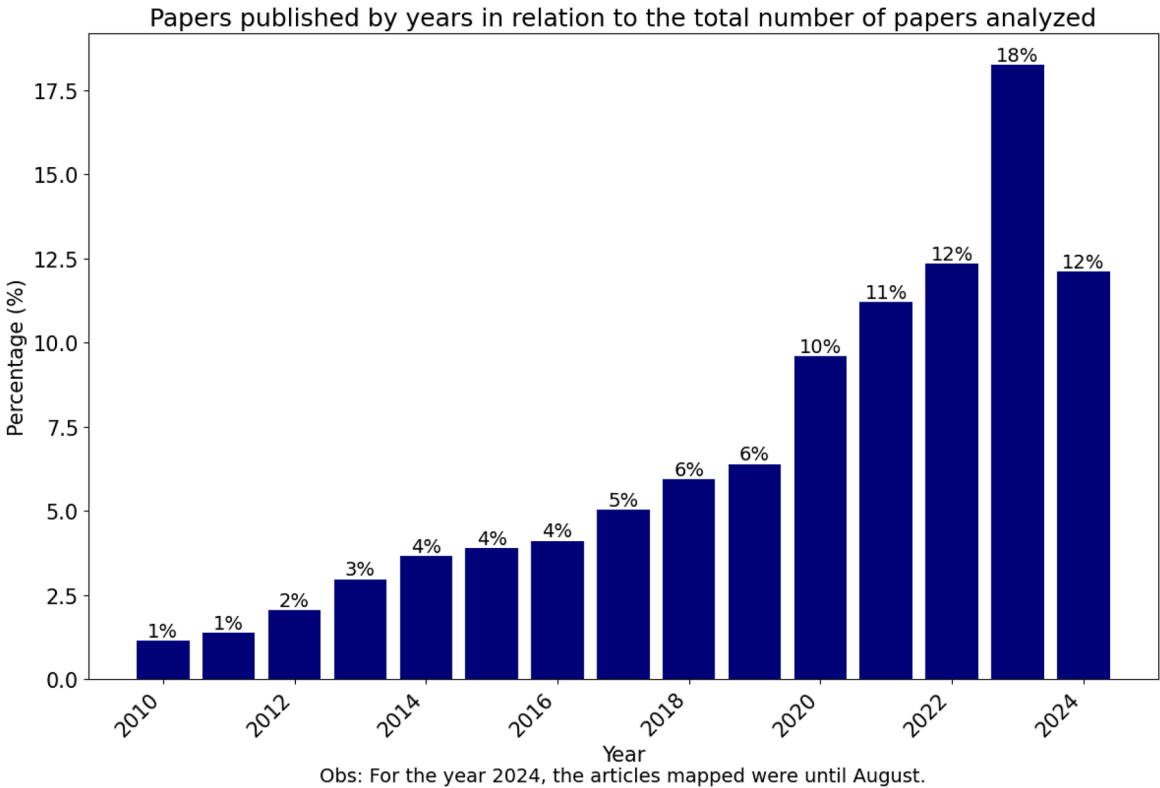


Fig. 11. Number of publications over time.

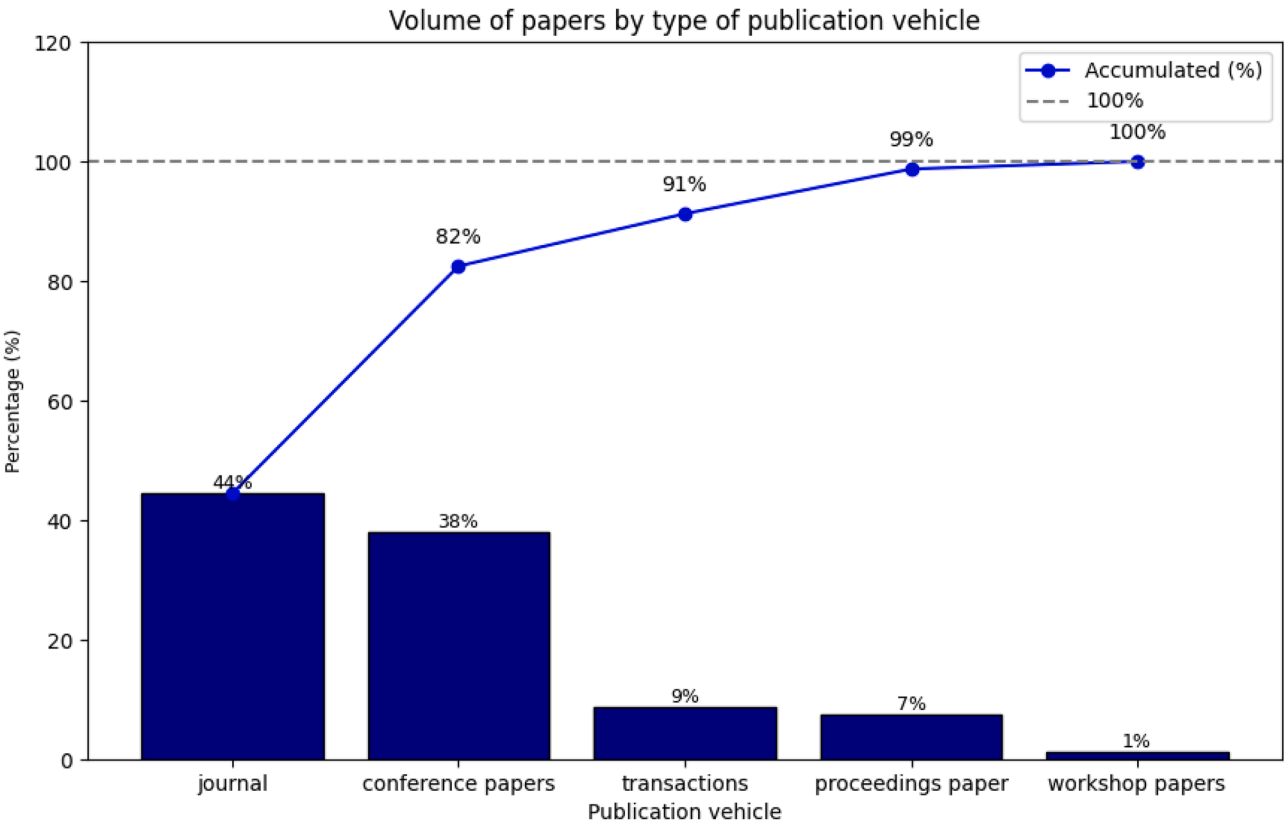
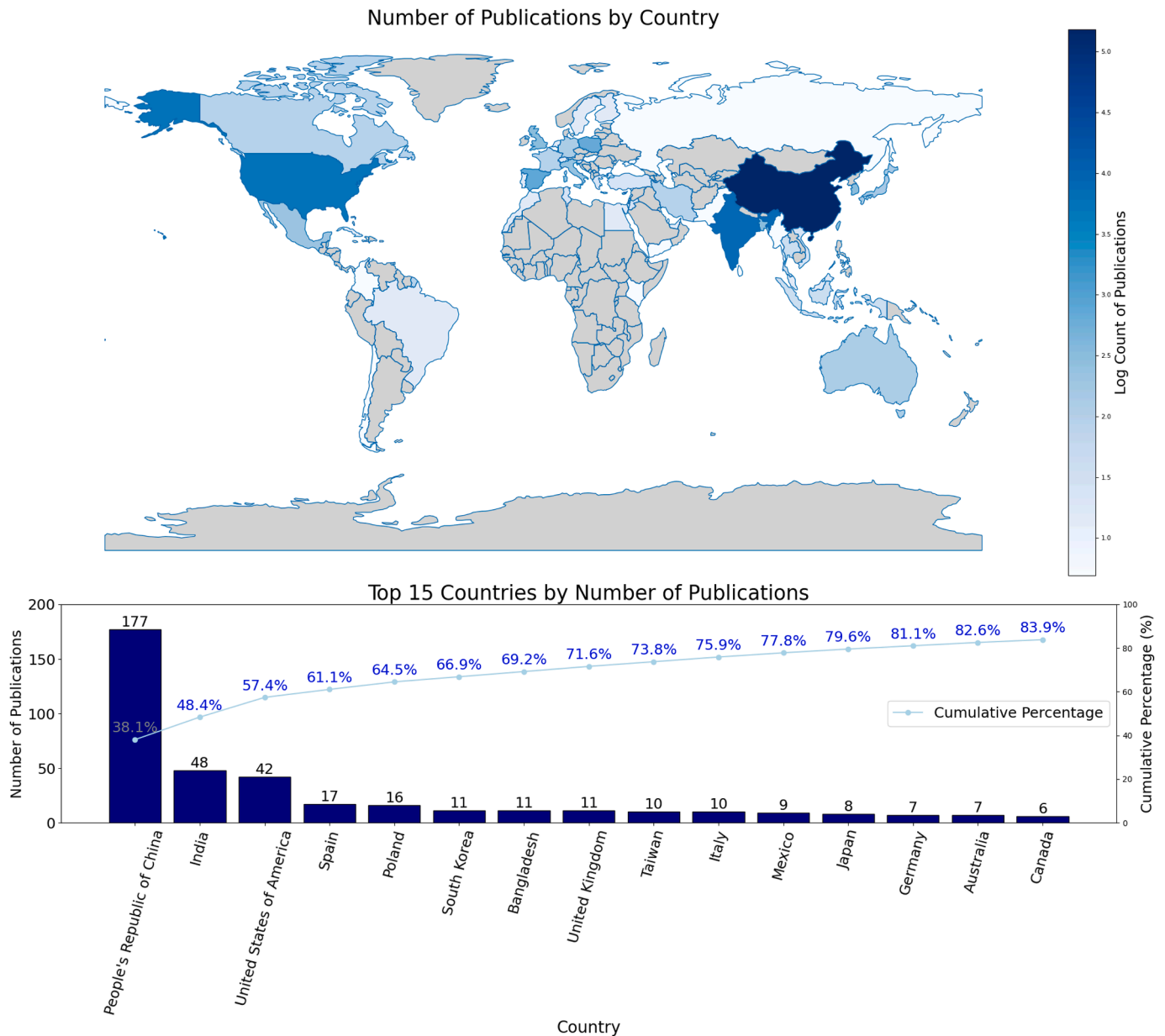


Fig. 12. Number of publications by type.



**Fig. 13.** Distribution of papers produced by country.

classifier performance but also result in biased models with low generalization capacity, negatively impacting predictions and yielding inaccurate results (Fig. 17).

**7.2. RQ2: What techniques and methodologies have been most commonly used to address class imbalance, and what are their advantages and disadvantages?**

From a general perspective, the results confirm that data sampling-based methods remain the dominant approaches for addressing class imbalance, aligning with findings from other systematic reviews (Hairani et al., 2024; Susan & Kumar, 2021).

By categorizing the methodologies into the three classes previously discussed in the preliminaries section (Sneha & Annappa, 2024), we find that 61 % of the approaches employ hybrid methods (Fig. 18).

Applying a grouping that examines multimodal approaches, we observe that regarding the type of data used, there is a significant predominance of studies focused on tabular data, with approximately 84 % of the

analyzed works employing this format (Fig. 19). This trend emphasizes the need to investigate methods that address multimodal data, which poses a considerable challenge in the field.

Using the classification (the same classification/grouping used in the bibliometric analysis), considering the 468 selected articles, we identified a total of 777 classifications, highlighting that most studies combine multiple methodologies to address the class imbalance problem. The methodology is strongly influenced by the type of data analyzed. For image data, we observed a predominance of Deep Learning-based approaches, accounting for 41 % of the methodologies used for this type of data. For tabular data, the most prevalent category was Sampling Theory, representing 31 % of the proposed methodologies. The distribution of methodologies employed across different data types is represented in Fig. 20.

For multimodal data, distinct trends emerge. In the tabular-image domain, the predominant methods were Cost-Sensitive and Sampling Theory, each accounting for 26 % of occurrences. Meanwhile, for text-image data, Deep Learning was the most commonly used approach,

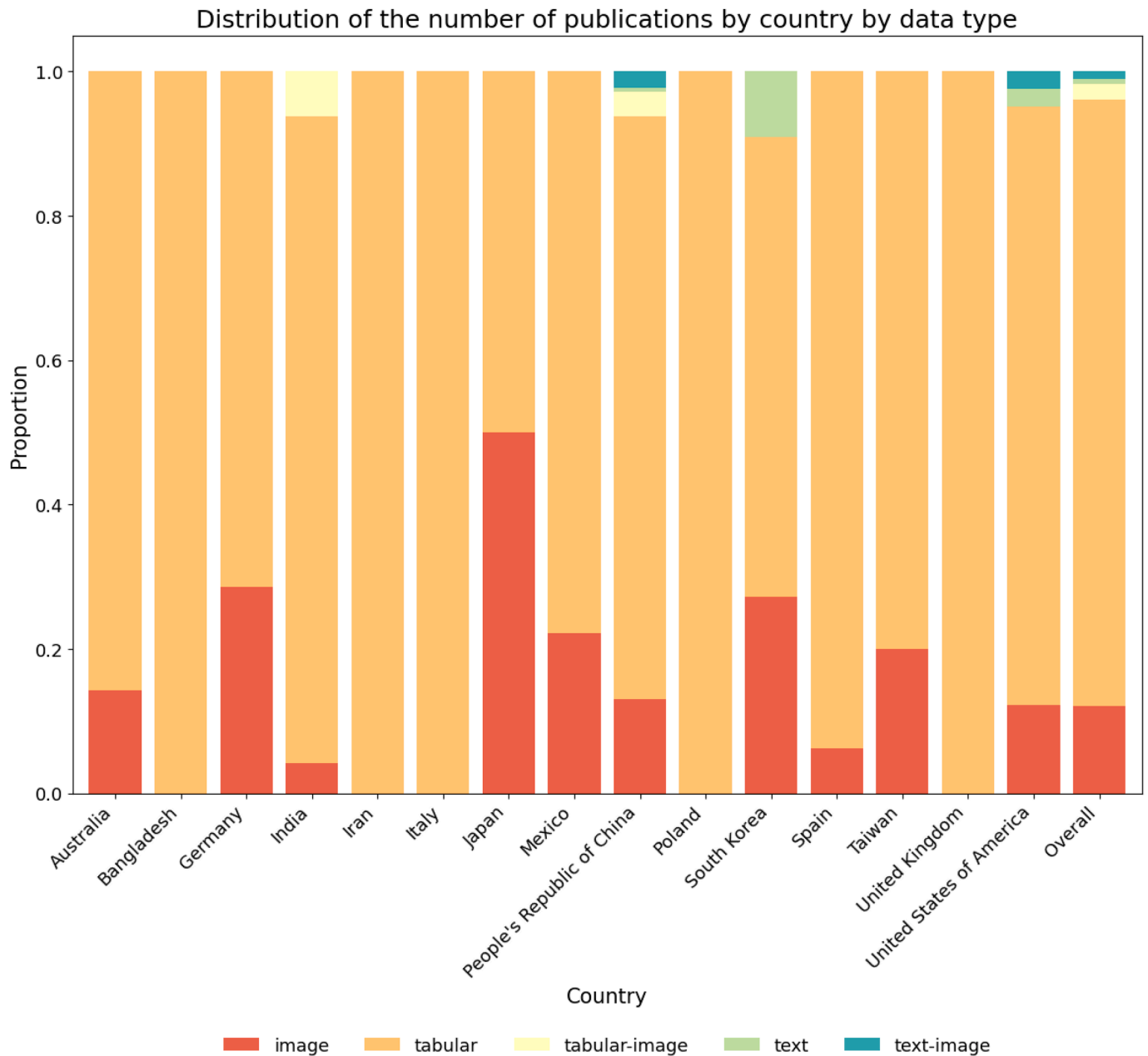


Fig. 14. Volume of publications by data type.

representing 29 % of the identified methodologies. Fig. 21 highlights the intersections between different data types and the proposed methodologies.

Regarding evaluation metrics, the analysis indicates that ROC-AUC (Fawcett, 2006), precision, F1-score (Murphy, 2012), MCC, PR-AUC, recall, and geometric mean (G-mean) (Sneha & Annappa, 2024) are commonly used, in line with the findings of the systematic review by Sneha and Annappa (2024). However, rather than debating metrics individually, we recommend selecting them by *category*, as operationalized in Section 8.1. Fig. 22 quantifies these findings.

**7.3. RQ3: In which application domains are these techniques most commonly applied?**

Data analysis of selected articles reveals that the majority (55 %) utilize datasets from diverse application domains (Fig. 23), followed by studies focused specifically on healthcare, finance, and security. This

diversity of data sources underscores the importance of the imbalance problem, which is evident across a wide range of societal sectors. The recurring presence of this challenge in different contexts highlights the urgent need for effective approaches to mitigate it, underscoring the importance of research that considers the specificities of each domain. Thus, it is crucial that future studies explore not only existing techniques, but also innovations that can be adapted to these varied scenarios.

**7.4. RQ4: What are the emerging trends and future research directions in the field of class imbalance in machine learning?**

Emerging trends and future research directions in the field of class imbalance emphasize the importance of validating proposed methodologies on real-world data, allowing for a more in-depth analysis of approaches in applied real contexts. Furthermore, the emergence of hybrid methods that integrate various techniques (both at the data and



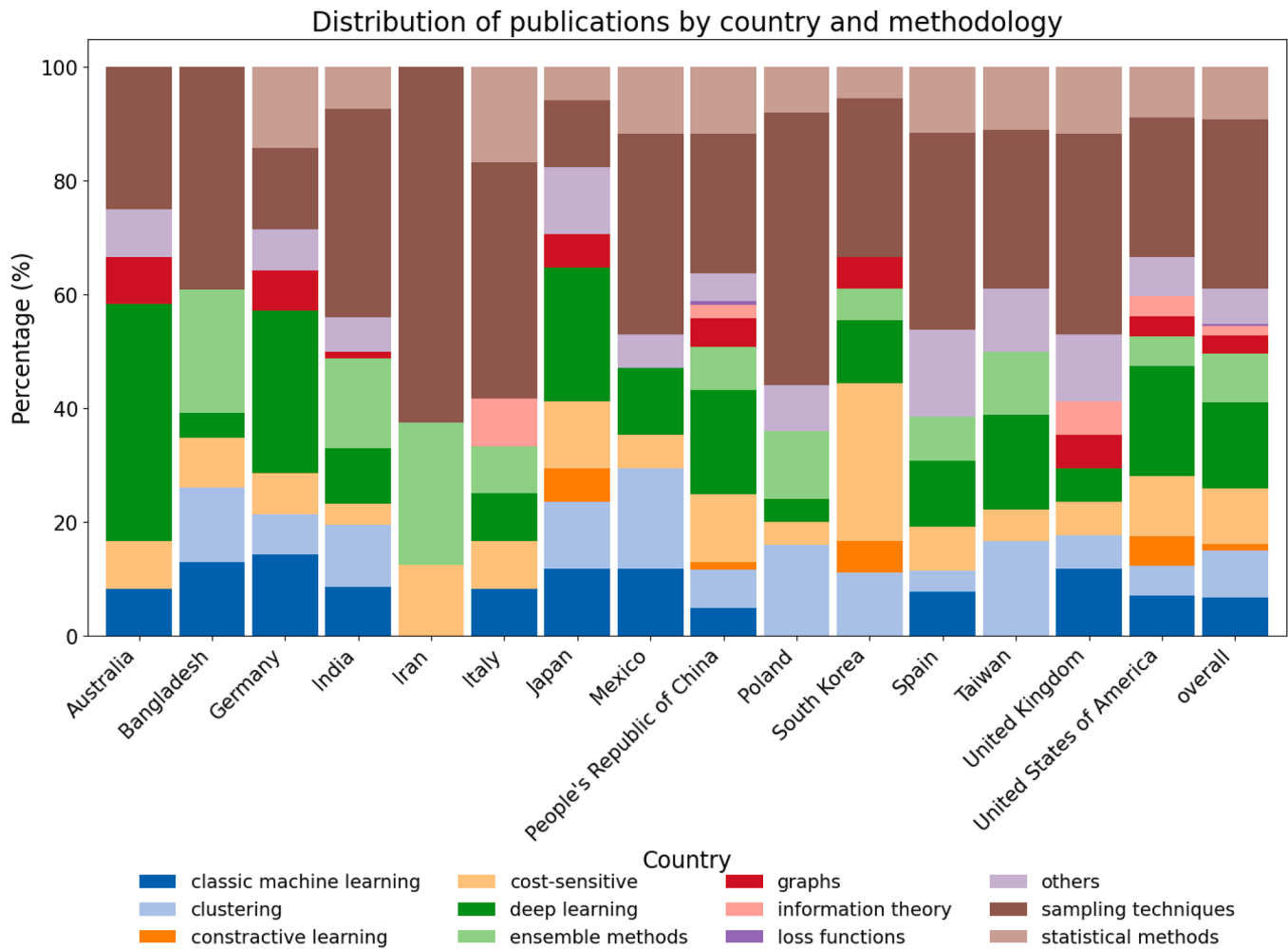


Fig. 15. Volume of publications by methodology.

algorithmic levels) has been identified as a promising solution to address the challenges associated with imbalance at all stages of model training, validation, and application. Imbalanced learning in multiclass tasks, while representing a significant challenge, offers tremendous potential for application across various domains.

Using the grouping method presented in Research Question 2 (RQ2), we analyzed the evolution of the publication volume for each methodology by data type over the past ten years. We chose to exclude the year 2024 from this analysis, as full access to publications for the year is not yet available, ensuring greater accuracy in interpreting the results. Additionally, a more detailed analysis of emerging trends will be presented later.

From a consolidated perspective, it is observed that sampling-based methods have predominated over the past decade, being widely employed to address class imbalance. However, an analysis of recent trends reveals a significant increase in the volume of publications related to deep learning-based approaches, suggesting that this category may become dominant in the literature in the coming years. Other methods that have shown continuous growth over the analyzed period include cost-sensitive approaches and graph-based methods (Fig. 24).

When analyzing data types specifically, we found that the structure of the data directly influences the choice of approach for handling class imbalance (Fig. 25). In the case of image data, a predominance of deep learning-based methods is observed, reflecting the efficiency of Convolutional Neural Networks (CNNs) and related architectures in automatic feature extraction. In contrast, for tabular data, sampling-based

methods remain the most frequently adopted approach, indicating that techniques such as SMOTE and undersampling are still widely used to balance class distributions in this domain.

With the advancement of deep learning (LeCun et al., 2015), the treatment of data imbalance has been directly incorporated into neural network architectures, either through the design of specific modules or through modifications to loss functions. Thus, an emerging trend in contemporary research can be observed: Following significant advancements enabled by data manipulation methods such as SMOTE and its variations, the focus has shifted toward the evolution of strategies that integrate the treatment of imbalance into the very architectures of neural networks, promoting more robust and integrated solutions. In 2024, among the 48 selected articles, approximately 21 (44% of the total in 2024) were related to methods based on deep neural networks.

Regarding the 21 articles related to deep learning, 43% focus on the use of Convolutional Neural Networks (CNNs), 24% on Graph Neural Networks (GNNs), and 14% on Generative Adversarial Networks (GANs). Together, these architectures account for 81% of the publications on deep learning in the context of class imbalance in 2024, as illustrated in Fig. 26.

When considering the centralization of learning (in terms of how data is stored and processed for training machine learning models), among the 48 analyzed articles, 5 focus on federated learning, while the remaining ones adopt the traditional approach. Notably, among the 5 papers on federated learning in the context of class imbalance, 4 were published in 2024, indicating that this is an emerging research topic

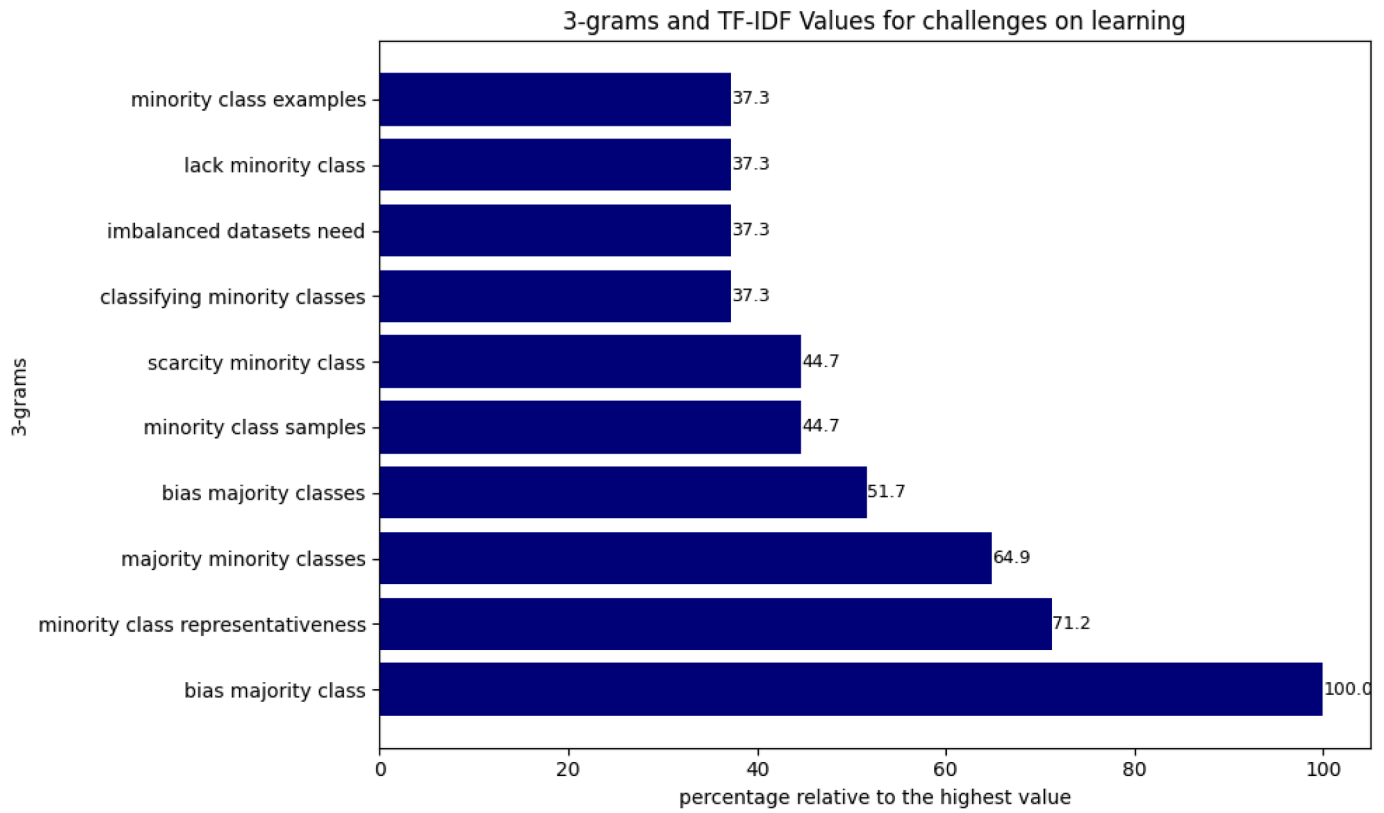


Fig. 16. RQ1 - Challenges faced in imbalanced learning.

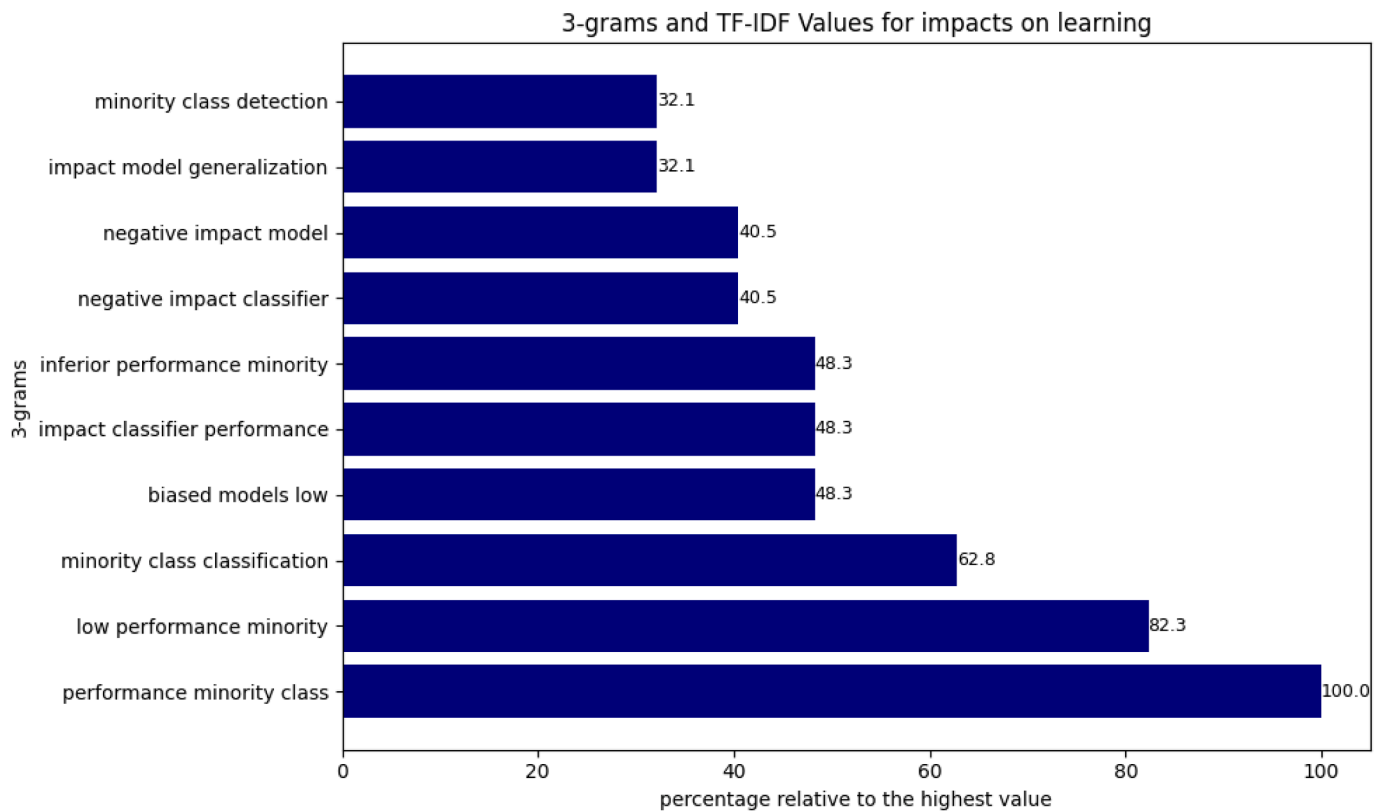


Fig. 17. RQ1 - Impacts Caused in the Process.

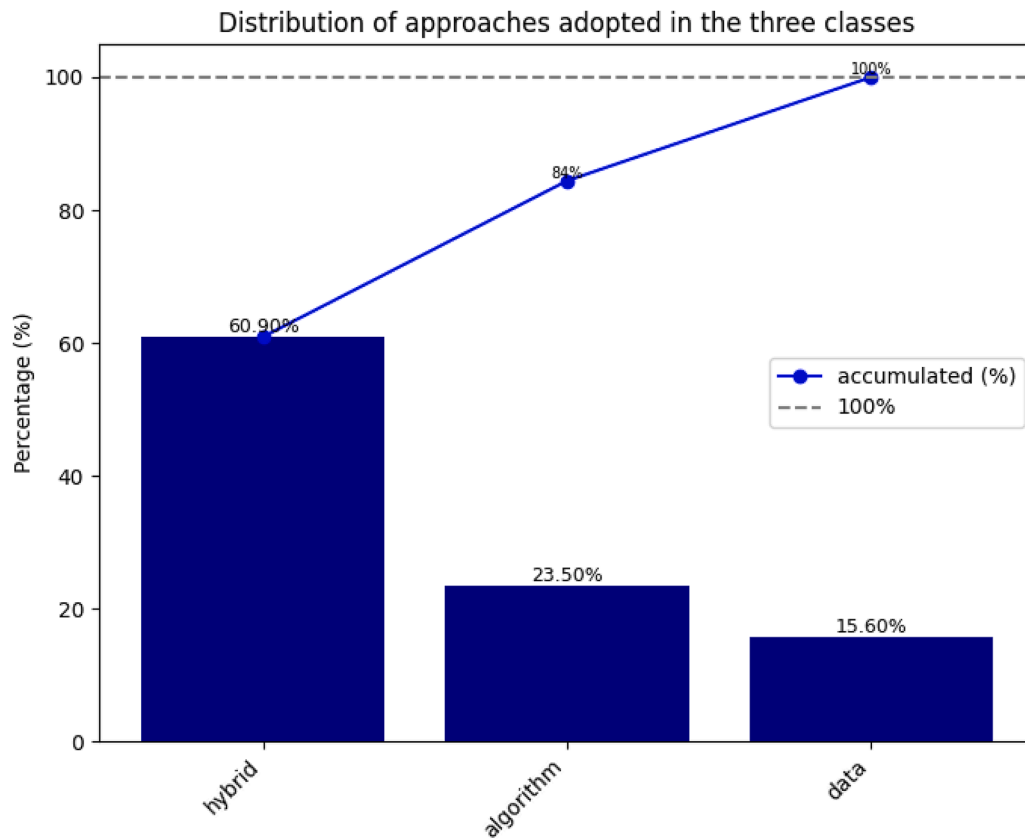


Fig. 18. RQ2 - Clusters of the most commonly used techniques.

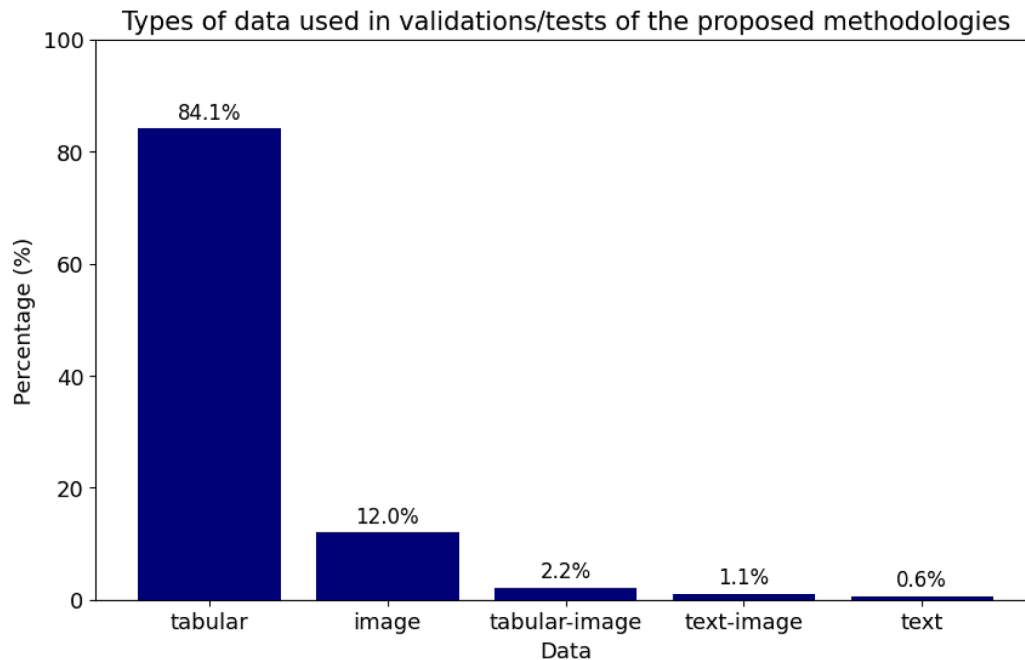


Fig. 19. RQ2 - Data type.

(Chen & Shen, 2024). With the advancement of AI applications across various sectors, federated learning is expected to play a crucial role in the development of scalable and optimized solutions.

Furthermore, contrastive learning and incremental learning have been gaining increasing attention, driven primarily by the maturation

of machine learning models for commercial-scale applications. A comprehensive overview of these topics is presented below.

#### 7.4.1. Convolutional neural networks (CNNs)

Convolutional Neural Networks (CNNs) have become one of the leading approaches for deep learning in various fields, primarily due to their

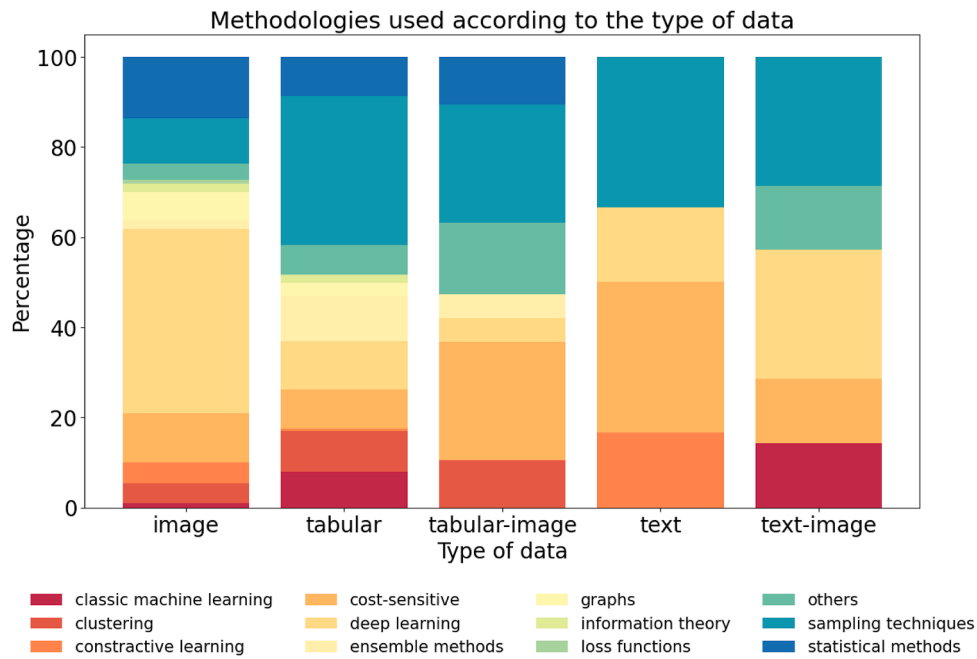


Fig. 20. RQ2 - Use of methodologies according to the type of data.

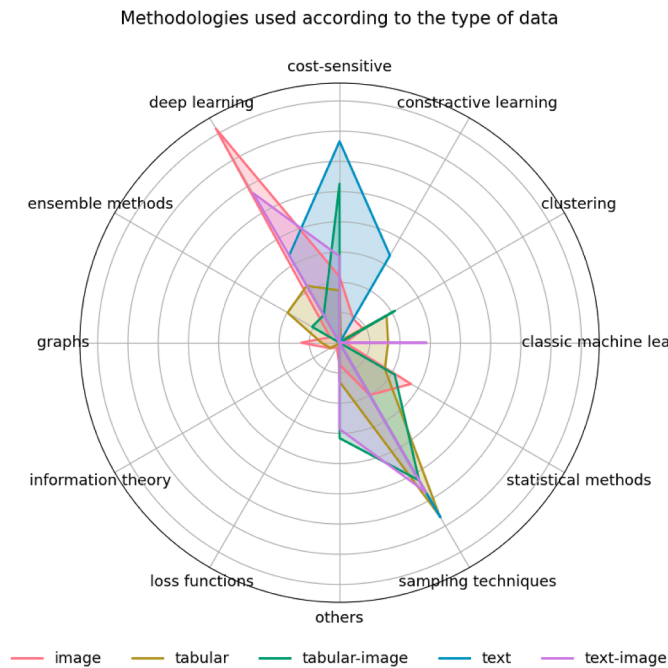


Fig. 21. RQ2 - Main methodologies by type of data.

ability to automatically extract relevant features from structured data, such as images and signals. However, their effectiveness can be severely impacted by class imbalance in training data. To mitigate this challenge, various approaches have been proposed, including class reweighting, sampling techniques, and specialized architectures.

In Table 2, we present a synthesis of key findings from nine recent articles that explore the use of CNNs in the context of learning with class imbalance.

The analyzed studies demonstrate significant advancements in applying CNNs for learning with class imbalance. Models employing class reweighting have shown improved balance between majority and minority classes, while hybrid architectures integrating CNNs with Transformers have proven promising for handling severe imbalance

conditions. Incremental learning strategies also hold potential for enhancing adaptation in dynamic data scenarios.

#### 7.4.2. Generative adversarial networks (GANs)

Generative Adversarial Networks (GANs) were introduced by Goodfellow et al. (2014) as a method for generating synthetic data through competition between a generator and a discriminator. While the generator creates artificial samples, the discriminator attempts to distinguish them from real ones, iteratively improving the quality of the generated samples.

Originally designed for image synthesis, GANs quickly gained prominence in various machine learning domains due to their ability to model complex distributions without requiring explicit probability functions

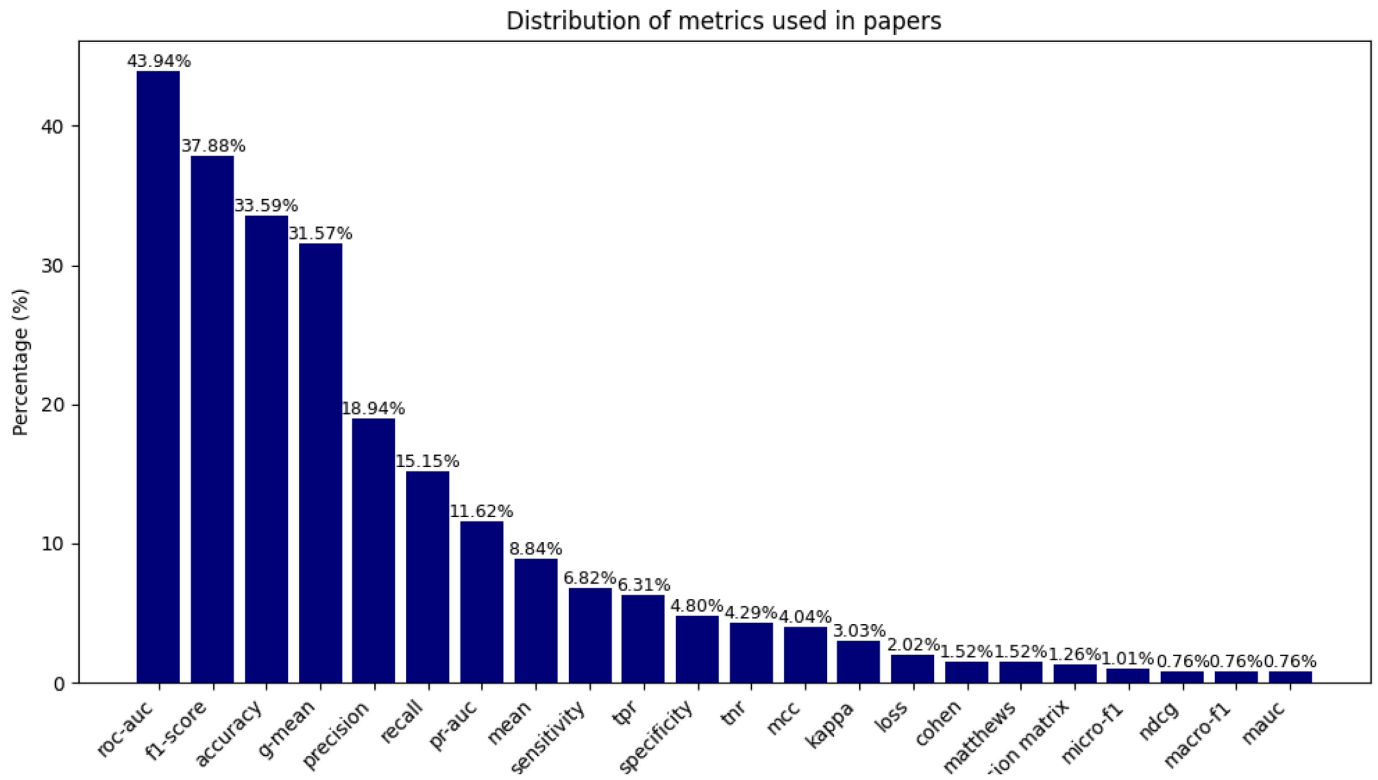


Fig. 22. RQ2 - Most commonly used metrics.

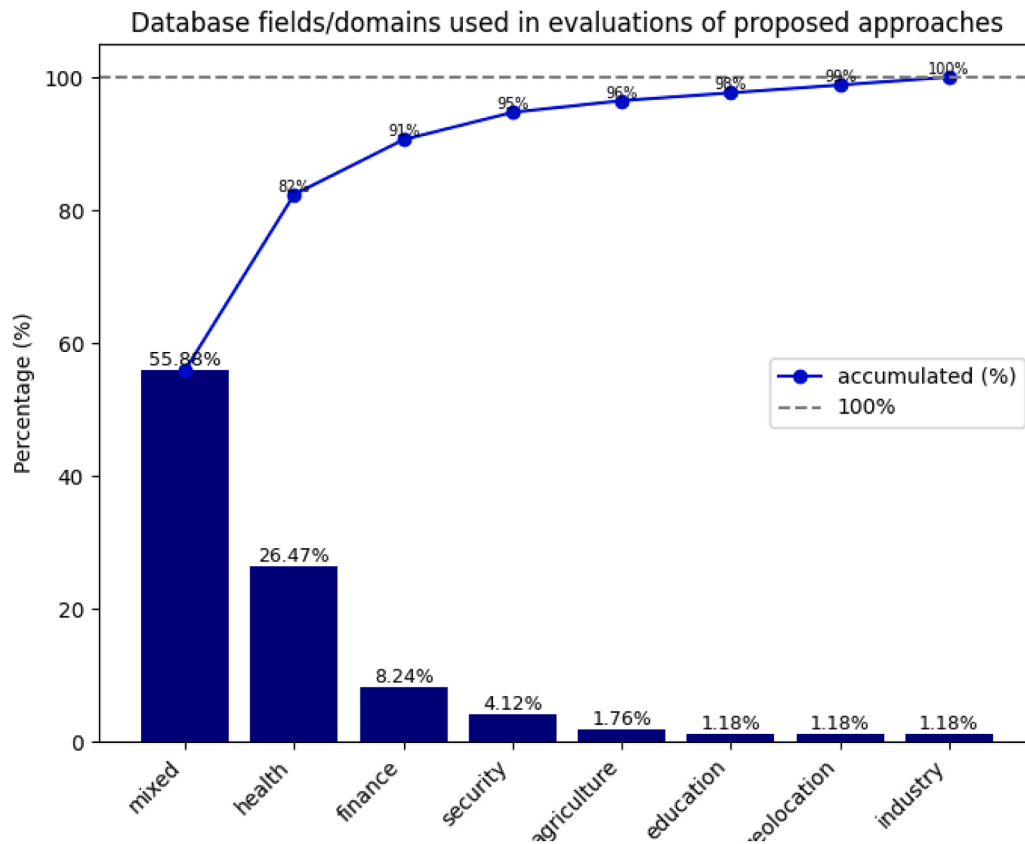
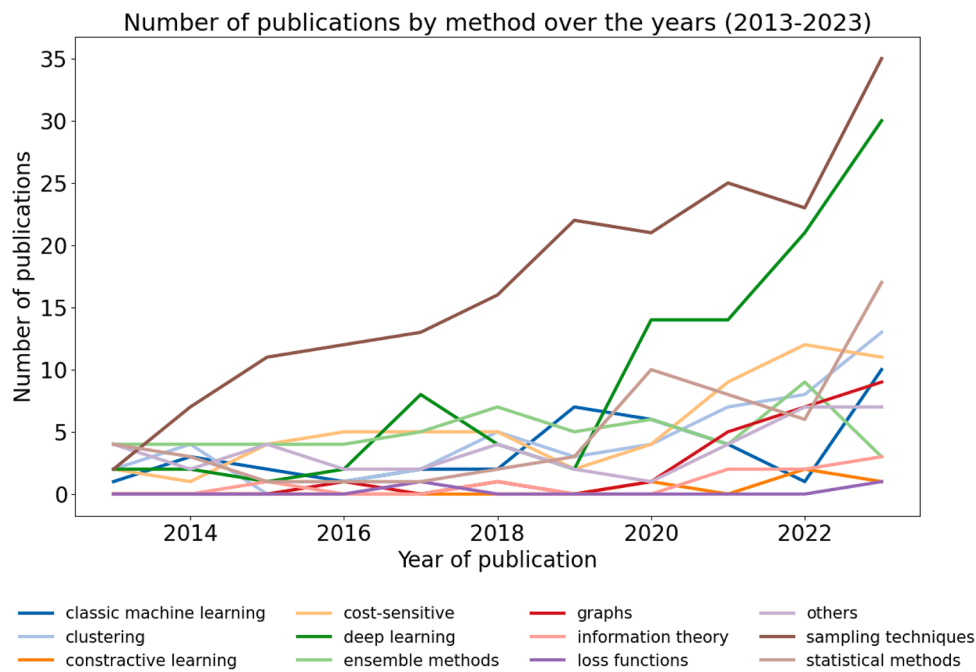


Fig. 23. RQ3 - Subject/Domain of the data to which the proposed technique was applied.





**Fig. 24.** Evolution of methodologies over the years.

Table 2

### Synthesis of key findings and validation of recent studies on CNNs for learning with class imbalance.

Article	Key Findings	Validation
No One Left Behind: Real-World Federated Class-Incremental Learning (Dong et al., 2023)	Proposes an anti-forgetting model in Federated Learning to handle class imbalance among clients. Introduces adaptive loss functions for gradient compensation and semantic distillation to mitigate knowledge loss.	Experiments on non-IID federated datasets, including CIFAR-100 and Tiny-ImageNet.
Otem-IGCD: An Optimal Transport-based EM Framework for Imbalanced Generalized Category Discovery (Li et al., 2024d)	Develops an optimal transport-based method for category discovery in imbalanced data, ensuring better distribution between known and unknown classes.	Evaluated on CIFAR-100 and ImageNet-100.
Rolling Bearing Intelligent Fault Diagnosis Towards Variable Speed and Imbalanced Samples Using Multiscale Dynamic Supervised Contrast Learning (Dong et al., 2024)	Introduces a multi-scale CNN with an attention mechanism to identify faults in bearings under severe imbalance conditions. Uses dynamic supervised contrast learning to enhance class differentiation.	Tested on two bearing fault datasets with a 20:1 imbalance ratio.
TCP: Triplet Contrastive-Relationship Preserving for Class-Incremental Learning (Li et al., 2024b)	Proposes a triplet-based contrastive approach to preserve relationships between old and new classes in incremental learning, addressing challenges posed by class imbalance.	Experiments with CIFAR-100 and ImageNet under reduced memory constraints.
An Adaptive Bagging Algorithm Based on Lightweight Transformer for Multi-Class Imbalance Recognition (Wang et al., 2024)	Develops a hybrid model combining CNNs and lightweight Transformers with adaptive bagging techniques to improve classification accuracy in imbalanced data.	Validated on CIFAR-10-LT, CIFAR-100-LT, and Places365-LT.
DACA: A Domain Adaptive Fault Diagnosis Approach with Class-Aware Based on Cross-Domain Extreme Imbalance Data (Li et al., 2024c)	Proposes a domain adaptation method for fault diagnosis in environments with extreme imbalance, using CNNs and adversarial learning for domain alignment.	Tested in three extreme imbalance protocols with mechanical fault datasets.
DOC3: Deep One-Class Classification Using Contradictions (Dhar & Gonzalez-Torres, 2024)	Introduces a contradiction-based one-class classification approach to enhance anomaly detection in highly imbalanced data scenarios.	Evaluation on tabular and image datasets, including an anomaly detection benchmark.
IOSL: Incremental Open Set Learning (Ma et al., 2023)	Develops an incremental learning model to handle the continuous evolution of classes and detect new categories in imbalanced scenarios.	Experiments on incremental datasets, including MNIST and CIFAR-100.
Multimodal Imbalanced Data Fault Diagnosis Method Based on a Dual-Branch Interactive Fusion Network (He et al., 2024)	Introduces a multimodal fusion architecture based on CNNs for fault diagnosis, incorporating interactive learning to mitigate class imbalance effects.	Tested on multimodal industrial fault datasets.

(Goodfellow et al., 2014). In class-imbalanced machine learning, they have emerged as a promising strategy to enhance the representation of minority classes and improve predictive performance (Luo et al., 2023; Mariani et al., 2018).

In imitation learning, where agents learn from demonstrations, class imbalance can lead to mode collapse, favoring only the most frequent

behaviors (Gu & Zhu, 2024). To mitigate this issue, Gu and Zhu (2024) proposed Balanced *Generative Adversarial Imitation Learning* (BAGAIL), which employs GANs to balance the distribution of samples. The discriminator is trained to distinguish not only between real and synthetic samples but also between different behavioral modes, ensuring representative diversity. BAGAIL is trained in two stages: (i) a pre-training

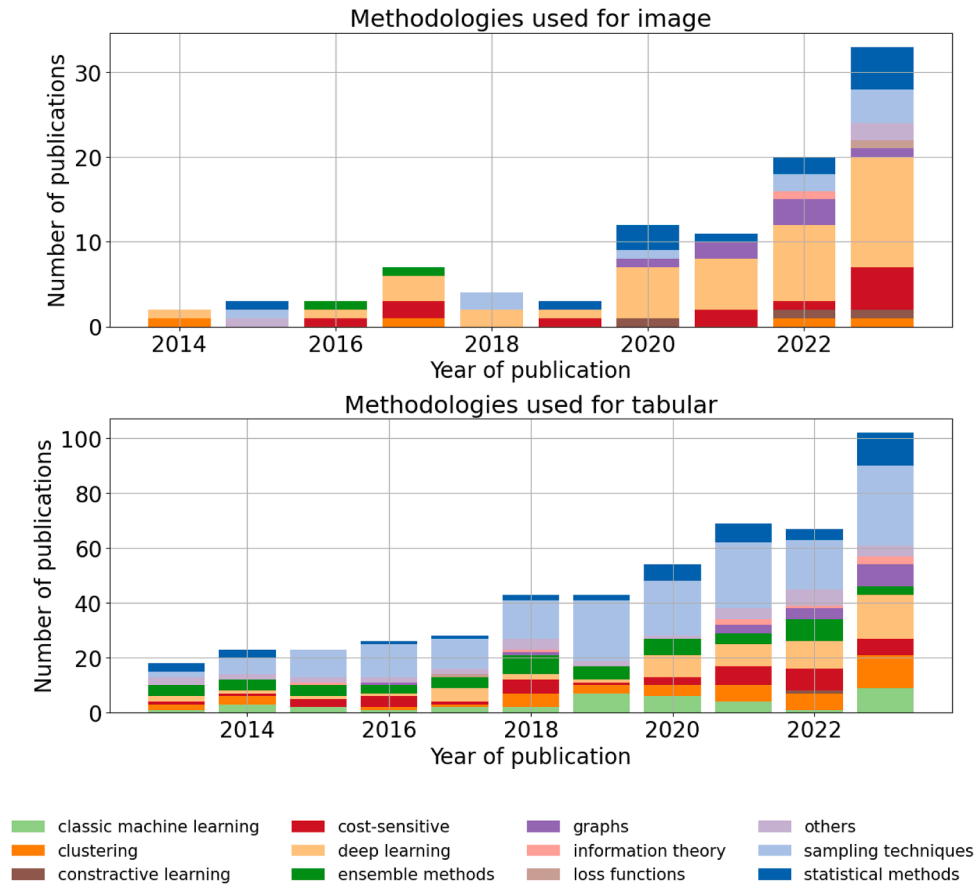


Fig. 25. Methodology by data type (tabular and image only).

phase stabilizes the generator and accelerates convergence, followed by (ii) adversarial refinement, where the generator improves its trajectories and the discriminator enhances its distinction capability. Experiments demonstrated that BAGAIL outperformed GAIL and ACGAIL in imitation learning scenarios, ensuring better representation of diverse behavioral styles.

Another challenge in imbalanced learning is anomaly detection, where extreme imbalance can bias classifiers towards the majority class, hindering the identification of critical failures. To address this issue, Bougaham et al. (2024) developed *Vector Quantized GAN Anomaly Detection Through Intermediate Patches* (VQGanoDIP), which combines Vector Quantized GAN (VQGAN) with localized anomaly detection techniques. VQGanoDIP operates in three stages: (i) training VQGAN with normal images, (ii) reconstructing input images, and (iii) analyzing differences between original and reconstructed versions. The final binary classifier adjusts detection to minimize false negatives (ZFN). Tests showed that VQGanoDIP achieved 94.65 % accuracy on the MVTec-AD dataset and 87.93 % on a private automotive dataset, demonstrating its effectiveness in detecting subtle anomalies.

For tabular data, Das (2024) proposed a hybrid approach combining Association Rules with *Tabular GAN* (TGAN) to balance imbalanced datasets in supervised classification. The technique reduces the majority class by removing redundant samples and generates new synthetic samples for the minority class, mitigating biases common in conventional resampling methods and improving classifier generalization.

Experiments demonstrated that this approach is promising compared to SMOTE and ADASYN, proving effective in extreme imbalance scenarios. The combination of informed redundant sample removal and high-quality synthetic data generation reduced false negatives and improved model robustness.

In summary, GANs have proven effective in handling class imbalance across various scenarios, from imitation learning to anomaly detection

and tabular classification. However, challenges such as dynamic hyperparameter tuning, high computational cost, and difficulty in preserving diversity and representativity in synthetic samples still require significant advancements. Emerging approaches, including enhanced conditional GANs, active learning, meta-learning, and generative diffusion model integration, represent promising paths toward making these techniques more robust, efficient, and generalizable in highly imbalanced and complex environments.

#### 7.4.3. Graph neural networks (GNNs)

*Graph Neural Networks* (GNNs) have emerged as a powerful tool for learning on graphs, enabling the capture of complex relationships between connected entities. The key concept behind these networks is message passing, where each node updates its representation based on its neighbors. This principle was introduced by Scarselli et al. (2008) in 2008 through recursive neural networks.

Since then, several advancements have been made in the field. Methods such as the *Graph Convolutional Network* (GCN) (Kipf & Welling, 2016) improved this process by applying convolutional filters directly to the graph structure. Another relevant approach was GraphSAGE (Hamilton et al., 2017), which introduced sampling techniques to enhance model scalability.

Despite these advances, most GNNs assume that data is homogeneous and balanced, which is rarely the case in real-world applications. Class imbalance represents a significant challenge, as message passing tends to favor majority classes, thereby hindering the representation of minority classes (Van Belle & De Weerd, 2024; Xia et al., 2024; Xu et al., 2024).

In Table 3, we present a synthesis of recent works that explore the use of GNNs to address class imbalance, highlighting the proposed techniques and their experimental validations:

The analyzed studies show important advancements in adapting GNNs to imbalanced learning. Techniques like pre-aggregation

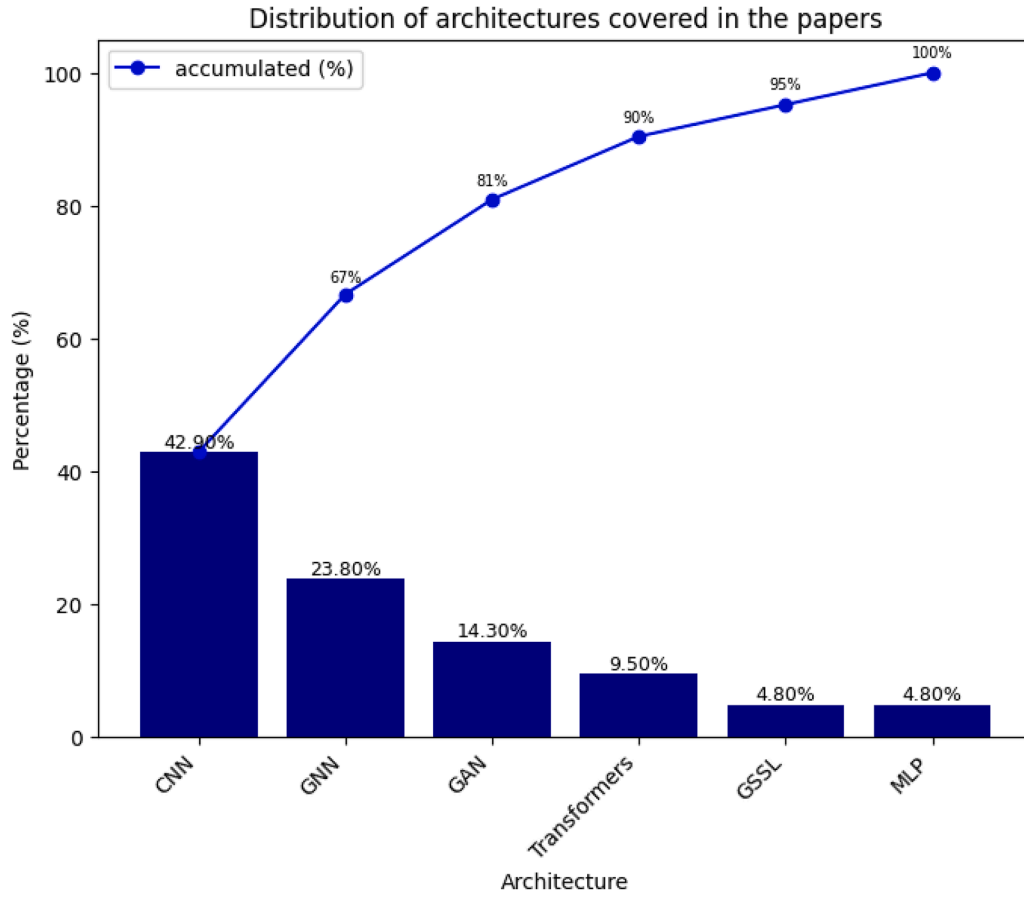


Fig. 26. Neural network architectures addressed in the papers.

Table 3

Summary of recent studies addressing class imbalance using Graph Neural Networks.

Article	Key Findings	Validation
A novel graph oversampling framework for node classification in class-imbalanced graphs (Xia et al., 2024)	Proposed the <i>Distribution Alignment-based Oversampling</i> (Graph-DAO) method for node classification with class imbalance, introducing distribution alignment using SPNs and pre-aggregation representations via autoencoders.	Outperformed GraphSMOTE and GraphMixup in classification accuracy on benchmark datasets.
Revisiting graph-based fraud detection in sight of heterophily and spectrum (Xu et al., 2024)	Introduced the <i>Spectrum-Enhanced and Environment-Constrained Graph Fraud Detector</i> (SEC-GFD), which combines spectral filtering and an environmental constraint module to improve fraud detection in graphs with heterophily and imbalance.	Demonstrated better performance than GraphSAGE and GAT in highly imbalanced and heterophilic networks.
SHINE: A Scalable Heterogeneous Inductive Graph Neural Network for Large Imbalanced Datasets (Van Belle & De Weerd, 2024)	Developed the <i>Scalable Heterogeneous Inductive Graph Neural Network</i> (SHINE), a scalable heterogeneous GNN that addresses imbalance via noise reduction, heterogeneous aggregation, and inductive convolutional layers.	Achieved improved scalability and accuracy compared to GraphSAGE and R-GCN in dynamic graph settings.

oversampling, spectral filtering, and heterogeneous aggregation have proven effective in mitigating issues such as majority class dominance and heterophily. Nevertheless, future research is still needed to develop unified frameworks that can handle class imbalance, dynamic topologies, and domain generalization simultaneously, particularly in real-world contexts such as fraud detection, biomedical graph analysis, and recommendation systems.

#### 7.4.4. Contrastive and incremental learning methods

Contrastive learning emerged as a method for constructing representations by comparing pairs of samples (Becker & Hinton, 1992). Initially conceived as a technique to capture invariances in data, it was gradually expanded to various applications within machine learning. During the 1990s, supervised learning still heavily relied on large volumes of labeled data, which encouraged the development of techniques

**Table 4**  
Scenario-based user cases aligned with the taxonomy.

Scenario	Regime	Topology	Algorithms	Examples	Evaluation Categories (focus)
Tabular, moderate IR, binary or multiclass	Batch	Centralized	Hybrid (oversampling + cost-sensitive / class weights; threshold tuning)	SMOTE and variations	Threshold-free + Operating-point
Tabular, extreme IR or label noise	Batch	Centralized	Hybrid with noise handling	Clustered / Borderline-SMOTE + ENN/Tomek; cost-sensitive loss; threshold moving by cost	Threshold-free + Operating-point
Image, multiclass	Batch	Centralized	Algorithm-level	CNN; GAN, MLP	Threshold-free + Operating-point; Aggregation & slicing
Text, multilabel or multiclass	Batch	Centralized	Hybrid	PU/weak supervision + class weights; contrastive pretraining; per-label thresholds	Threshold-free + Operating-point; Aggregation & slicing
Graphs	Batch or streaming	Centralized	Hybrid	Hybrids: Graph-DAO, GraphSAGE, SHINE	Threshold-free + Operating-point

capable of modeling representations without the explicit need for labels. One of the first concrete examples of its practical application was presented by Bromley et al. (1993), who introduced Siamese Networks for handwritten signature verification, demonstrating the potential of contrastive learning in real-world scenarios.

In the context of supervised learning, methods such as *Supervised Contrastive Learning* (SCL) (Khosla et al., 2020) have demonstrated significant advantages over traditional loss functions, such as cross-entropy. More recently, Khalid et al. (2024) proposed a variation of supervised contrastive learning in which label embeddings are used as anchors to better organize the data distribution in the latent space. This approach, called *Label Supervised Contrastive Learning* (LSCL), proved effective in text classification tasks with imbalanced data, contributing to better class separation and reducing biases induced by training distribution. Furthermore, the authors explored the application of adapted spaces, such as hyperbolic spaces, to hierarchically structure representations, which proved particularly useful in scenarios where class relationships play a central role. Indeed, hyperbolic spaces are advantageous as they efficiently capture hierarchies and sparse distributions, facilitating the organization of imbalanced classes in the latent space.

On the other hand, in Class-Incremental Learning (CIL), the imbalance between old and new classes introduces additional challenges, such as catastrophic forgetting (French, 1999). To address this issue, Li et al. (2024b) presented the *Triplet Contrastive Preserving* (TCP) method, a contrastive learning-based approach that preserves structural relationships between samples. Unlike conventional approaches that apply point-to-point constraints to embeddings of old classes, TCP maintains contrastive relationships between old and new samples, allowing the model to incorporate new classes without compromising previously acquired knowledge. This strategy demonstrated significant performance gains in incremental classification tasks, especially in scenarios where memory constraints limit the storage of past samples. The ability of contrastive learning to preserve structural relationships between samples makes this approach particularly relevant for incremental learning, as it reduces the degradation of previous representations when integrating new classes.

Additionally, the study also proposed *Asymmetric Augmented Contrastive Learning* (A2CL), a mechanism that adjusts transformations applied to old and new samples to minimize bias in incremental learning. The results indicated that introducing asymmetry in data augmentation enhances the model's ability to learn new representations without degrading performance on previously learned classes. Methods like this indicate a promising path toward the convergence of contrastive and incremental learning, enabling models to evolve continuously without losing valuable information.

Recent advances make it clear that contrastive learning plays a fundamental role in mitigating class imbalance, both in static and incremental scenarios. Approaches such as LSCL and TCP highlight the importance of structuring the embedding space in a way that preserves class separation and maintains knowledge retention over time. However, several challenges remain, particularly regarding the efficient integration of these approaches with large-scale deep learning architectures.

Issues such as the optimization of computational efficiency in contrastive training, the adaptation of incremental learning to large-scale dynamic data, and the development of hybrid methods that combine supervised contrastive learning, resampling strategies, and adaptive mechanisms remain open. Advancing in these directions could lead to the development of more resilient and generalizable models, capable of handling imbalanced and constantly evolving contexts.

#### 7.4.5. Federated learning

Federated Learning, introduced by Google in 2017, enables machine learning models to be trained in a decentralized manner, ensuring data privacy by keeping data within local devices/environments. Only model updates are sent to a central server for aggregation. Unlike conventional approaches that require data centralization, this technique minimizes privacy risks but faces significant challenges in scenarios involving imbalanced and non-IID (Non-Independent and Identically Distributed) data.

In many practical cases, local environments exhibit highly heterogeneous data distributions, reflecting different user behavioral patterns. This variation results in biased local models, hindering the aggregation of a cohesive global model. Consequently, majority classes tend to be favored, which compromises performance on minority classes, particularly in critical applications such as medical diagnosis and security.

To address this issue, Li et al. (2024a) propose a hybrid approach based on an adjusted loss function and prototype sharing among clients. The loss function statistically balances local predictions, reducing the impact of class imbalance. Prototype sharing fosters consistency in local model representations, improving convergence stability. In experiments conducted on CIFAR-10, CIFAR-100, and Tiny-ImageNet, this technique demonstrated up to a 9.5% increase in per-class accuracy compared to FedAvg.

In semi-supervised federated learning, Chen and Shen (Chen & Shen, 2024) address the underutilization of unlabeled data, a common limitation in traditional approaches. To overcome this problem, they propose the *Exploitation Maximization for Federated Semi-Supervised Learning* (EM-FSSL) framework, which integrates *Entropy Meaning Loss* (EML) and *Adaptive Negative Learning* (ANL). While EML adjusts the prediction distribution to prevent distortions, ANL utilizes adaptive negative

pseudo-labels, optimizing the generalization of the global model. The FullMatch framework, developed from these techniques, outperformed FixMatch, achieving up to a 7.3% accuracy improvement on CIFAR-100 and enhancements in minority class classification.

In supervised federated learning, Zhu et al. (2024) introduce *Federated Simulated Centralized Learning* (FedSCL), inspired by centralized learning, where data aggregation naturally reduces imbalance. The method combines *Serial Updating*, which sequentially transfers the global model across clients, and *Parallel Strategy*, which distributes multiple copies of the model for simultaneous training. This approach accelerates convergence and improves model stability, validated on MNIST, CIFAR-10, and Tiny-ImageNet-200, surpassing FedAvg, especially in minority class performance.

In the context of incremental learning on stream data, where new data is continuously introduced, Dong et al. (2023) investigate catastrophic forgetting, which affects the retention of old classes. To mitigate this issue, they propose *Local-Global Anti-forgetting (LGA)*, which combines dynamic gradient adjustments, semantic distillation, and prototype communication at the server level. Experiments on CIFAR-100 and ImageNet demonstrated up to a 12.7% performance improvement in extreme imbalance scenarios, ensuring more stable continuous learning.

Some key topics that could be explored in future research include:

- **Privacy and regulatory aspects:** Some strategies, such as prototype sharing, may reveal sensitive patterns. Model sharing must be carefully designed, especially in imbalanced cases where a particular class may be favored and, therefore, uniquely identifiable.
- **Scalability and computational efficiency:** Centralized aggregation can create communication bottlenecks in networks with thousands of clients. Future directions involve research into decentralized and parallelized methods.
- **Generalization across domains:** This remains a challenge, as models such as FedSCL may require substantial adjustments when applied to new contexts. Meta-learning strategies and knowledge transfer could enhance the adaptability of these models.
- **Model interpretability:** Finally, interpretability is essential for critical applications such as healthcare and security. Research in explainability for federated learning in the context of class imbalance could contribute to the adoption of such models in critical environments such as healthcare.

## 8. Synthesis: Empirically derived taxonomy and practical guidelines

### 8.1. An empirically derived taxonomy

Motivated by the evidence consolidated in RQ2, e.g., hybrid approaches in 61% of studies, tabular data in 84%, and deep learning for image tasks in 41%, we derive an empirically grounded taxonomy that connects problem context, solution algorithms, and evaluation. Although taxonomies for imbalanced learning exist (Sneha & Annappa, 2024), they do not take the context of the problem in consideration, as well the data modality and other factors that are critical when applying such algorithms.

The taxonomy operationalizes "what to use, when, and how to evaluate" by closing the loop from context to algorithms and to evaluation, reflecting the patterns observed in our mapping (Figs. 18–20 and 22). In practice, regime and topology can be decomposed into concrete scenarios (e.g., batch-centralized, batch-federated, incremental-centralized), which we use in the selection guideline table.

To operationalize evaluation in imbalanced settings, we group metrics into five categories. The categories are complementary—not mutually exclusive—and examples are provided only for orientation; specific choices should follow the application's operating point, costs, and regulatory constraints.

- **Threshold-free / ranking-based (global):** assess the quality of the ranking across all possible cut-offs; no single decision threshold is fixed. Suited for global model comparison under severe imbalance and for retrieval-type goals (e.g., PR-AUC / Average Precision for minority-centric retrieval; ROC-AUC for ranking discrimination).
- **Threshold-dependent / operating-point:** quantify performance at a chosen decision threshold (or at top- $k$ ). Appropriate when a fixed operating point is mandated by clinical/operational policy or when deployment requires a concrete cut-off (e.g., per-class recall/precision, F1, balanced accuracy, confusion-matrix rates; top- $k$  for vision tasks).
- **Calibration / probabilistic quality:** measure how well predicted probabilities match observed frequencies. Critical whenever decisions are cost-sensitive or when thresholds are set from probabilities (e.g., Brier score, log-loss, ECE/MCE; calibration curves).
- **Cost-/utility-aware decisioning:** incorporate asymmetric error costs or clinical/net-benefit trade-offs directly into evaluation and threshold selection (e.g., expected cost/utility, cost curves, decision-curve analysis; thresholds chosen by cost).
- **Aggregation & slicing protocol:** specify how results are aggregated and where they must be disaggregated. Report the aggregation scheme (macro vs. micro vs. weighted) and slice results per class; per client (federated learning); head/medium/tail (long-tailed regimes); per phase (incremental/streaming); and per label and example-based (multilabel).

**Positioning.** Unlike prior taxonomies that primarily organize methods by family or imbalance mechanism (Sneha & Annappa, 2024), our scheme explicitly binds *problem context* (task, modality, regime, topology, data difficulties) to *solution algorithms* and then to *evaluation categories*. This closes the loop from "when/where" to "what/how to use and how to report", and serves as a direct scaffold for the scenario-based guideline in Section 8.2. In this sense, the taxonomy is not only descriptive but also *operational*, turning the mapping's evidence into concrete decision paths.

### 8.2. Practical guidelines and user cases

This closing section provides a *minimal, scenario-based guideline* grounded on the taxonomy in Fig. 27. It is intentionally concise and non-prescriptive: the **application domain**, **operating constraints**, and **data difficulties** ultimately determine the final pipeline. Rather than enumerating all method combinations, we recommend leveraging the taxonomy to *structure* decisions and the literature to *specialize* them to the target context.

*How to use this guide (focused steps).*

- **Taxonomy first (problem framing).** Identify *Task type* (binary, multiclass, multilabel, long-tailed), *Data modality/structure*, *Learning regime* (batch, streaming, incremental), *Data topology* (centralized, federated; non-IID severity), and *Data difficulties* (IR, overlap, small disjuncts, label noise, drift). This maximizes the use of information about the modeled problem and anchors all subsequent choices.
- **Literature triage by context.** Query surveys and empirical studies *matching your axes* (e.g., "long-tailed image classification balanced softmax", "multilabel text class imbalance per-label thresholding", "federated non-IID class imbalance prototypes"). Prior approaches that are *recurrent in your specific context*; this avoids reinventing the wheel and reduces the risk of overfitting decision choices.
- **Algorithms (first-line baseline, then escalate).** Start with the *first-line baseline* for the scenario and escalate complexity only if validation plateaus or deployment constraints require it. Use *data-level* → *algorithm-level* → *hybrid/domain-specific* as the escalation path.
- **Evaluation by metric categories.** Pair *threshold-free/ranking* (global comparison) with *threshold-dependent/operating-point* at the required



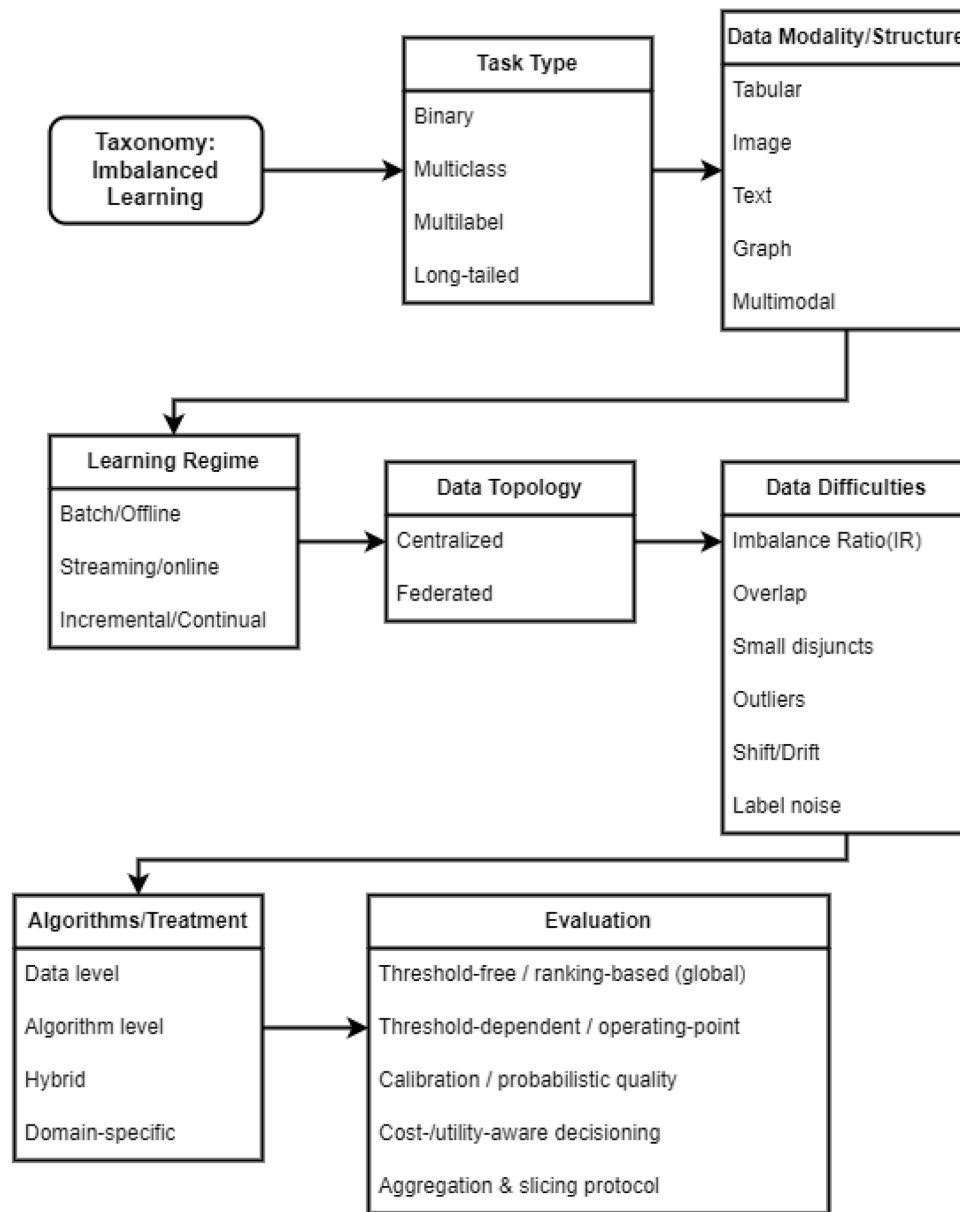


Fig. 27. Empirically derived taxonomy for imbalanced learning.

operating point; add *calibration* when probabilities drive decisions; use *cost-/utility-aware* analysis when cost/benefit policy is in place; and always declare the *aggregation & slicing* protocol (macro/micro/weighted; per-class; head-tail; per-client; per-phase; per-label).

- **Document and decide.** Record objectives, cost matrices, data splits (avoiding leakage), final thresholds, and stability checks (drift, client-level reporting, head-tail slicing). The final decision should be *traceable to the context* and *justified* by the evaluation categories.

To ensure comparability, reproducibility, and alignment with our taxonomy and evaluation framework, we recommend reporting at least the following:

#### Minimal reporting checklist..

- Declare context axes (task/modality/regime/topology/difficulties) and chosen algorithms.
- State IRs and data difficulties (overlap, small disjuncts, noise) and how addressed.

- Use a set of metrics to evaluate the models—preferably from different categories 8.1, apply data windowing, and analyze the metrics average behavior.
- Share seeds, splits (no leakage), costs, and calibration diagnostics.

## 9. Conclusion

This systematic mapping aimed to provide a broad and structured overview of the current landscape of machine learning in class imbalance scenarios. The adopted strategy combined manual and automated techniques, leveraging machine learning models, semantic embeddings, and graph-based methods. The analysis involved an extensive literature search, with the screening of over 25953 articles, resulting in an in-depth analysis of 468 primary studies.

Based on the proposed research questions, it was possible to draw an up-to-date portrait of the field and identify promising directions for future investigations. The analysis of RQ1 highlighted the predominance of hybrid approaches, which combine data-level and algorithm-level imbalance treatments. These strategies, present in 61 %

of the studies, have shown consistent results precisely because they integrate multiple mechanisms for addressing imbalance, enhancing their overall effectiveness.

With regard to data types, RQ2 revealed a clear concentration on tabular data – used in 84 % of the studies –while modalities such as images and texts remain underexplored. This finding reveals a significant gap, especially considering that applications in computer vision and natural language processing deal with imbalances that have distinct and often more complex characteristics.

Concerning evaluation metrics, ROC-AUC emerged as the most frequently used, followed by F1-score and accuracy. However, the recurrent use of accuracy in imbalanced settings warrants attention: it tends to inflate results by prioritizing performance on the majority class, which may distort the model's actual ability to identify minority class examples – the very class that, in many cases, carries greater decision-making value. This reinforces the need for careful metric selection, particularly in asymmetric scenarios.

RQ4 shows that data sampling methods still dominate the publications in the field, but also emphasizes that deep learning approaches are rapidly gaining ground, suggesting they may become the dominant topic in the coming years. Among the emerging areas, we highlight the use of CNNs for visual tasks, graph-based representations –which are well-suited for handling irregular data – and federated learning, which stands out for combining decentralization and privacy preservation, an essential feature in sensitive domains such as healthcare and finance. Other promising directions include self-supervised learning and contrastive learning, both of which offer more adaptive solutions, especially in settings with limited labeled data.

Still within the scope of RQ4, the mapped challenges go beyond technical issues. There are also methodological and regulatory barriers that must be considered. One critical point is scalability: many models are still validated only in controlled environments with small datasets. Model explainability also emerges as a central concern, particularly in regulated domains, where the use of deep networks can hinder the interpretability of results. In addition, legal and ethical requirements – especially in sectors such as law and healthcare-demand transparency and algorithmic accountability. Added to this is the complexity of working with multimodal data – such as images, text, and physiological signals – which still poses a major technical challenge when it comes to integrating and balancing diverse input modalities.

To consolidate the practical contributions of this study, we developed an empirically derived taxonomy that explicitly links the problem-context axes (task type, data modality, learning regime, topology, and data difficulties) to the solution levers (data-level, algorithm-level, hybrid, or domain-specific techniques) and to the corresponding evaluation categories. Building on this framework, we provide a minimal, scenario-based guideline (4) that states “what to use, when, and how to evaluate,” complemented by an essential reporting checklist. Together, the taxonomy and the guideline translate the mapped landscape into reproducible decision paths, facilitating the coherent selection of techniques and metrics in real-world machine-learning projects with class imbalance.

In summary, this study provides not only an accurate depiction of the current landscape but also a strategic starting point for future research. The combination of a solid selection methodology and a detailed taxonomy of the analyzed approaches strengthens the reliability of the findings. It is expected that this work will serve as a foundation for more targeted research initiatives and foster the development of fairer, more effective, and practically applicable solutions in real-world scenarios where class imbalance remains a significant barrier in machine learning.

#### Declaration of generative AI in scientific writing

Statement: During the preparation of this work the author(s) used chatGPT 4.0 in order to Assistance with translation. After using this

tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

#### CRedit authorship contribution statement

**Gilberto Sussumu Hida:** Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Visualization, Writing - original draft; **André Câmara Alves Do Nascimento:** Supervision, Writing - review & editing.

#### Data availability

Data will be made available on request.

#### 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.

#### Supplementary material

Supplementary material associated with this article can be found in the online version at [10.1016/j.eswa.2025.129592](https://doi.org/10.1016/j.eswa.2025.129592)

#### References

- Becker, S., & Hinton, G. E. (1992). Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356), 161–163.
- Bougaham, A., El Adoui, M., Linden, I., & Frénay, B. (2024). Composite score for anomaly detection in imbalanced real-world industrial dataset. *Machine Learning*, 113(7), 4381–4406.
- Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1993). Signature verification using a "siamese" time delay neural network. *Advances in Neural Information Processing Systems*, 6(6), 737–744.
- Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Mercer, R. L., & Roossin, P. (1988). A statistical approach to language translation. In *Coling budapest 1988 vol 1: International conference on computational linguistics* (pp. 1–6).
- 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.
- 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.
- Chen, S., & Shen, J. (2024). Exploitation maximization of unlabeled data for federated semi-supervised learning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, (pp. 1–6). <https://doi.org/10.1109/TETCI.2024.3361866>
- Das, S. (2024). A new technique for classification method with imbalanced training data. *International Journal of Information Technology*, 16(4), 2177–2185.
- de Moraes, R. F., Miranda, P. B. C., & Silva, R. M. A. (2016). A meta-learning method to select under-sampling algorithms for imbalanced data sets. In *2016 5th Brazilian conference on intelligent systems (BRACIS)* (pp. 385–390). IEEE.
- Dhar, S., & Gonzalez-Torres, B. (2024). Doc 3: Deep one class classification using contradictions. *Machine Learning*, 113(8), 5109–5150.
- Dong, J., Li, H., Cong, Y., Sun, G., Zhang, Y., & Van Gool, L. (2023). No one left behind: Real-world federated class-incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4), 2054–2070.
- Dong, Y., Jiang, H., Yao, R., Mu, M., & Yang, Q. (2024). Rolling bearing intelligent fault diagnosis towards variable speed and imbalanced samples using multiscale dynamic supervised contrast learning. *Reliability Engineering & System Safety*, 243, 109805.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.
- Felix, E. A., & Lee, S. P. (2019). Systematic literature review of preprocessing techniques for imbalanced data. *IET Software*, 13(6), 479–496.
- Ferreira, G. F., Quiles, M. G., Nazaré, T. S., Rezende, S. O., & Demarzo, M. (2021). Automation of article selection process in systematic reviews through artificial neural network modeling and machine learning: Protocol for an article selection model. *JMIR Research Protocols*, 10(6), e26448.
- French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4), 128–135.
- Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., & Japkowicz, N. (2024). The class imbalance problem in deep learning. *Machine Learning*, 113(7), 4845–4901.
- Goodfellow, I., 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*, 27, 2672–2680.
- Gu, S., & Zhu, F. (2024). Bagail: Multi-modal imitation learning from imbalanced demonstrations. *Neural Networks*, 174, 106251.

- Hairani, H., Widiyaningtyas, T., & Prasetya, D. D. (2024). Addressing class imbalance of health data: A systematic literature review on modified synthetic minority oversampling technique (SMOTE) strategies. *JOIV: International Journal on Informatics Visualization*, 8(3), 1310–1318.
- Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, 30, 1024–1034.
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.
- He, J., Yin, L., & Sheng, Z. (2024). Multimodal imbalanced-data fault diagnosis method based on a dual-branch interactive fusion network. *IET Science, Measurement & Technology*, 18(7), 373–384.
- Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), 1–54.
- 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 (CSUR)*, 52(4), 1–36.
- Khalid, B., Dai, S., Taghavi, T., & Lee, S. (2024). Label supervised contrastive learning for imbalanced text classification in euclidean and hyperbolic embedding spaces. In *Proceedings of the ninth workshop on noisy and user-generated text (w-NUT 2024)* (pp. 58–67).
- Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33, 18661–18673.
- Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, .
- Kitchenham, B., Charters, S. et al. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report Technical report, ver. 2.3 EBSE technical report. ebse. 65.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Li, L., Zhan, D.-c., & Li, X.-c. (2024a). Aligning model outputs for class imbalanced non-IID federated learning. *Machine Learning*, 113(4), 1861–1884.
- Li, R., Sun, Y., & Wang, H. (2020). An imbalanced data classification method based on multi-kernel extreme learning machine fusion. In *2020 IEEE 3rd international conference of safe production and informatization (IICSPI)* (pp. 103–111). IEEE.
- Li, S., Ning, X., Zhang, S., Guo, L., Zhao, T., Yang, H., & Wang, Y. (2024b). TCP: Triplet contrastive-relationship preserving for class-incremental learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2031–2040).
- Li, Y., Zhu, Y., Yu, Y., Mao, R., Ye, L., Liu, Y., Liu, R., Lang, T., & Zhang, J. (2024c). Daca: A domain adaptive fault diagnosis approach with class-aware based on cross-domain extreme imbalance data. *Expert Systems with Applications*, 256, 124944.
- Li, Z., Dai, B., Meinel, C., & Yang, H. (2024d). Otem-IGCD: An optimal transport-based EM framework for imbalanced generalized category discovery. In *2024 International joint conference on neural networks (IJCNN)* (pp. 1–9). IEEE.
- Lipitakis, A.-D., & Lipitakis, E. A. (2014). On machine learning with imbalanced data and research quality evaluation methodologies. In *2014 International conference on computational science and computational intelligence* (pp. 451–457). IEEE (vol. 1).
- Luo, C., Xu, Y., Shao, Y., Wang, Z., Hu, J., Yuan, J., Liu, Y., Duan, M., Huang, L., & Zhou, F. (2023). Evagonet: An integrated network of variational autoencoder and wasserstein generative adversarial network with gradient penalty for binary classification tasks. *Information Sciences*, 629, 109–122.
- Ma, B., Cong, Y., & Ren, Y. (2023). IOSL: Incremental open set learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4), 2235–2248.
- Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., & Malossi, C. (2018). Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*, .
- Mikolov, T. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, .
- Murphy, K. P. (2012). Machine learning: A probabilistic perspective. MIT press.
- Ochal, M., Patacchiola, M., Vazquez, J., Storkey, A., & Wang, S. (2023). Few-shot learning with class imbalance. *IEEE Transactions on Artificial Intelligence*, 4(5), 1348–1358. <https://doi.org/10.1109/TAI.2023.3298303>
- Octaviano, F. R., Felizardo, K. R., Maldonado, J. C., & Fabbri, S. C. (2015). Semi-automatic selection of primary studies in systematic literature reviews: Is it reasonable? *Empirical Software Engineering*, 20, 1898–1917.
- Page, L. (1998). The pagerank citation ranking: Bringing order to the web. technical report. *Stanford Digital Library Technologies Project*, 1998, 17.
- Perez, E., Kiela, D., & Cho, K. (2021). True few-shot learning with language models. *Advances in Neural Information Processing Systems*, 34, 11054–11070.
- Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008). Systematic mapping studies in software engineering. In *12th International conference on evaluation and assessment in software engineering (EASE)* (pp. 68–77). BCS Learning & Development.
- Petersen, K., & Gerken, J. M. (2024). On the road to interactive LLM-based systematic mapping studies. *Information and Software Technology*, 178, 107611–107615.
- Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64, 1–18.
- Popoff, E., Besada, M., Jansen, J. P., Cope, S., & Kanter, S. (2020). Aligning text mining and machine learning algorithms with best practices for study selection in systematic literature reviews. *Systematic Reviews*, 9, 1–12.
- Provost, F. (2000). Machine learning from imbalanced data sets. In *Proceedings of the AAAI'2000 workshop on imbalanced data sets* (pp. 1–3). AAAI Press (vol. 68).
- Ragonesi, R., Morerio, P., & Murino, V. (2023). Learning unbiased classifiers from biased data with meta-learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1–9).
- Roth, S., & Wermer-Colan, A. (2023). Machine learning methods for systematic reviews: A rapid scoping review. *Delaware Journal of Public Health*, 9(4), 40.
- Saeidmehr, A., Steel, P. D. G., & Samavati, F. F. (2024). Systematic review using a spiral approach with machine learning. *Systematic Reviews*, 13(1), 32.
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, 10(3), e0118432.
- Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1), 61–80.
- Scherbakov, D., Hubig, N., Jansari, V., Bakumenko, A., & Lenert, L. A. (2024). The emergence of large language models (LLM) as a tool in literature reviews: an LLM automated systematic review. *arXiv preprint arXiv:2409.04600*, .
- Shakeel, F., Sabhitha, A. S., & Sharma, S. (2017). Exploratory review on class imbalance problem: An overview. In *2017 8th International conference on computing, communication and networking technologies (ICCCNT)* (pp. 1–8). IEEE.
- Shi, Y., Zhang, M., Zheng, X., & Chen, H. (2023). Class imbalanced semi-supervised learning with meta-learning. In *2023 International conference on cyber-physical social intelligence (ICCSI)* (pp. 186–190). IEEE.
- Sneha, H. R., & Annappa, B. (2024). Exploratory analysis of methods, techniques, and metrics to handle class imbalance problem. *Procedia Computer Science*, 235, 863–877.
- Susan, S., & Kumar, A. (2021). The balancing trick: Optimized sampling of imbalanced datasets—a brief survey of the recent state of the art. *Engineering Reports*, 3(4), e12298.
- Van Belle, R., & De Weerd, J. (2024). Shine: A scalable heterogeneous inductive graph neural network for large imbalanced datasets. *IEEE Transactions on Knowledge and Data Engineering*, 36(9), 4904–4915. <https://doi.org/10.1109/TKDE.2024.3381240>
- Van De Schoot, R., De Bruin, J., Schram, R., Zahedi, P., De Boer, J., Weijdem, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2), 125–133.
- Werner de Vargas, V., Schneider Aranda, J. A., dos Santos Costa, R., da Silva Pereira, P. R., & Victória Barbosa, J. L. (2023). Imbalanced data preprocessing techniques for machine learning: A systematic mapping study. *Knowledge and Information Systems*, 65(1), 31–57.
- Voudigari, E., Salamanos, N., Papageorgiou, T., & Yannakoudakis, E. J. (2016). Rank degree: An efficient algorithm for graph sampling. In *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 120–129). IEEE.
- Wang, J., Jiang, X., Liu, H., Cai, H., & Meng, Q. (2024). An adaptive bagging algorithm based on lightweight transformer for multi-class imbalance recognition. *Multimedia Systems*, 30(2), 99.
- Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of classification methods on unbalanced data sets. *IEEE Access*, 9, 64606–64628.
- Wu, O., & Li, M. (2024). Revisiting the effective number theory for imbalanced learning. *IEEE Transactions on Knowledge and Data Engineering*, 36(8), 4192–4206. <https://doi.org/10.1109/TKDE.2024.3367949>
- Xia, R., Zhang, C., Zhang, Y., Liu, X., & Yang, B. (2024). A novel graph oversampling framework for node classification in class-imbalanced graphs. *Science China Information Sciences*, 67(6), 1–16.
- Xu, F., Wang, N., Wu, H., Wen, X., Zhao, X., & Wan, H. (2024). Revisiting graph-based fraud detection in sight of heterophily and spectrum. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 9214–9222). (vol. 38).
- Zhang, H., & Ali Babar, M. (2010). On searching relevant studies in software engineering. In *14th International conference on evaluation and assessment in software engineering* (pp. 111–120). British Informatics Society Ltd.
- Zhang, H., Babar, M. A., Bai, X., Li, J., & Huang, L. (2011). An empirical assessment of a systematic search process for systematic reviews. In *15th Annual conference on evaluation and assessment in software engineering (EASE 2011)* (pp. 56–65). IET.
- Zhou, X., Wu, O., & Li, M. (2024). Investigating the sample weighting mechanism using an interpretable weighting framework. *IEEE Transactions on Knowledge and Data Engineering*, 36(5), 2041–2055. <https://doi.org/10.1109/TKDE.2023.3316168>
- Zhu, G., Liu, X., Niu, J., Wei, Y., Tang, S., & Zhang, J. (2024). Learning by imitating the classics: Mitigating class imbalance in federated learning via simulated centralized learning. *Expert Systems with Applications*, 255, 124755.