Multi-Class Imbalanced Data Classification: A Systematic Mapping Study

Yujiang Wang
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Malaysia
wyjprogram@gmail.com

Marshima Mohd Rosli
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Malaysia | Institute for Pathology, Laboratory and Forensic Medicine (I-PPerForM), University Teknologi MARA, Malaysia
marshima@uitm.edu.my (corresponding author)

Norzilah Musa
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Malaysia
norzi105@uitm.edu.my

Feng Li
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Malaysia | College of Computer and Information Engineering, Hebei Finance University, Baoding, China
chnlifeng2008@163.com

Received: 7 March 2024 | Revised: 2 April 2024 | Accepted: 4 April 2024

 DOI: https://doi.org/10.48084/etasr.7206

ABSTRACT

Multi-class data classification is distinguished as a significant and challenging research topic in contemporary machine learning, particularly when concerning imbalanced data sets. Hence, a thorough investigation of multi-class imbalanced data classification is becoming increasingly pertinent. In this paper, an overview of multi-class imbalanced data classification was generated via conducting a systematic mapping study, which endeavors to analyze the state of contemporary multi-class imbalanced data classification, with the primary goal of ascertaining the corpus of research undertaken in machine learning. To achieve this aim, 7,164 papers were assessed and the 147 prominent ones were selected from five digital libraries, which were further categorized according to techniques, issues, and types of datasets. After a thorough review of these papers, a taxonomy of multi-class imbalanced data classification techniques is proposed. Based on the results, researchers widely employ algorithmic-level, ensemble, and oversampling strategies to address the issue of multi-class imbalance in medical datasets, primarily to mitigate the impact of challenging data factors. This research highlights an urgent need for more studies on multi-class imbalanced data classification.

Keywords-multi-class imbalanced data; systematic mapping study; machine learning

I. INTRODUCTION

Imbalanced data classification has garnered widespread attention in the field of machine learning due to its relevance with real-world pattern classification tasks, where one or multiple classes (the minority) are significantly underrepresented as relative compared to the other classes (the majority) [1]. This situation is mimicked in countless real-world scenarios, such as fraud detection [2], disease diagnosis [3], and computer vision [4]. For example, regarding cancer malignancy grading, most patients are classified as healthy, while those with cancer are relatively rare. Consequently, it is becoming increasingly crucial to find ways to accurately identify and recognize patients with cancer to ensure adequate diagnosis and treatment. Thus, multi-class imbalanced data classification has drawn the interest of the machine learning community. The majority of imbalanced classification methods have explicitly focused on addressing skewed datasets in binary classification counterparts [5-8]. Several issues must be solved in multiclass imbalanced problems [7, 9]. Therefore, with this goal in mind, several class-imbalanced learning methods have been proposed. Typically, these can be divided into three primary categories: data-level, algorithm-level, and hybrid approaches [11]. Nevertheless, timely categorization requires further clarification. For instance, some articles define classification as a combination of oversampling and undersampling methods [5], while others describe it as combining data and algorithm-level techniques [11].

The design of a Systematic Mapping Study (SMS) provides a general overview of the research area and the published results [12]. In recent years, the SMS has gained widespread attention, leading to many studies incorporating it in classification tasks. In the medical domain, authors in [13] performed a systematic map of studies on pre-processing techniques in clinical datasets by grouping their publication years and channels, research, empirical, and contribution type. Authors in [14] conducted an SMS by identifying and analyzing empirical studies on ensemble classification methods for cardiological diseases published between 2000 and 2019. In total, 351 studies were selected and interpreted. Authors in [15] presented an SMS of how ensemble classifiers are exploited in the context of Intrusion Detection Systems (IDSs). A total of 124 prominent publications were collected, analyzed, and categorized based on the publication year, venues, datasets used, ensemble methods, and IDS techniques. The study fills a gap in the timely literature on an up-to-date systematic mapping study for IDSs. Moreover, significant scholarly interest has been focused on the research on SMS with ensemble classifiers. Alongside the above mentioned studies on ensemble classifiers, a generalized SMS executed in [16] considered 149 publications from the existing literature, which were gathered and examined. Conducting an SMS has also been adopted in various other domains, like fault prediction [17] and text classification [18], to gain deeper insight into the causes and nuances of classification problems.

However, as far as is known, SMS has not yet been performed on multi-class imbalanced data classification, underscoring the necessity for an overview mapping review to bridge this gap. Accordingly, this paper aims to provide a systematic mapping study of recent studies investigating multi-class imbalanced data classification. To accomplish this, 147 studies published from 2018 to 2023 were categorized and analyzed, while valuable insights into the most recent techniques, issues, and dataset types related to multi-class imbalanced data classification were acquired.

II. SYSTEMATIC MAPPING METHOD

In this section, an SMS is performed on multi-class imbalanced data classification, which is used to fill a knowledge gap in this area. Authors in [19] defined SMS as a methodology to build a classification scheme and quantify contributions concerning that classification. Although an SMS and a Systematic Literature Review (SLR) share similarities regarding the search process and study selection, their goals, data analysis, and evidence synthesis have significant differences. While SLR aims to synthesize timely evidence to generate fresh insights or conclusions, SMS is mainly concerned with structuring a research area without necessarily synthesizing proof [20]. Following the SMS definition in [19] an SMS for multi-class imbalanced data classification was conducted. Figure 1 shows the leading processes of this SMS. At first glance, it involves five stages: (1) identifying the research questions, (2) searching to gather studies, (3) screening the identified papers according to predetermined inclusion and exclusion criteria, (4) analyzing specifically selected papers, and (5) mapping the studies.

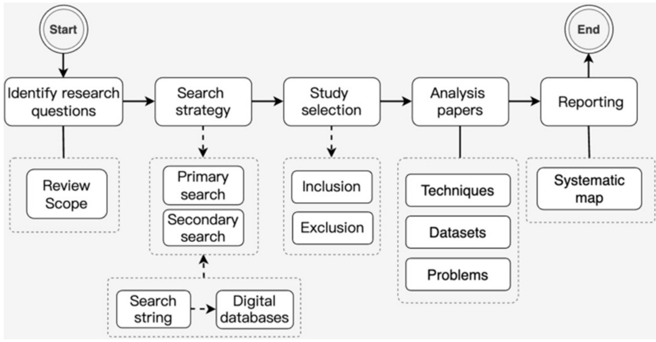

Fig. 1. The SMS process.

A. Research Questions

The present study's primary objective is to acquire an overview of the techniques, challenges, and widespread types of datasets involved in multi-class imbalanced data classification. With this in mind, three mapping Research Questions (RQs) have been identified, as illustrated in Table I, which serve as a foundation for constructing search strings for searches.

TABLE I. RESEARCH QUESTIONS

No.	Research question
RQ1	What techniques have been commonly used in multi-class imbalanced data classification?
RQ2	What are the primary issues when dealing with the classification of multi-class imbalanced data?
RQ3	What datasets are commonly utilised in multi-class imbalanced data classification studies?

B. Search Strategy

Primary studies were recognized through a search string in digital databases during this step. Subsequently, a meticulous manual search was conducted by individually browsing relevant conference proceedings and journal publications. This dual approach guarantees a comprehensive coverage, enabling the inclusion of high quality research papers [19]. For the collection of publications encompassing the fields of computer science from conference proceedings and journals, five reputable digital libraries, namely ScienceDirect, Scopus, IEEE Digital Library, ACM Digital Library, and Web of Science, were utilized. It is crucial to pre-select well-defined search keywords in order to obtain effective results in digital libraries. Hence, this study extracted keywords from the RQs and specified their alternative spellings and synonyms, based on those widely utilized in relevant publications. Specifically, the search process involved combining keywords using the Boolean operators "AND" and "OR". The first links the primary terms, while the latter incorporates synonyms. This fusion of keywords, as noticed in Table II, facilitates a targeted search approach. The "Scope" column displays the primary terms extracted on the basis of the RQs, whereas the "Search String" column presents additional synonyms associated with each term.

TABLE II. SEARCH STRING

Scope	Search String
multi-class	("multi-class" OR "multiclass" OR "multiple classes") AND
imbalanced	("imbalanced" OR "skewed" OR "unbalanced") AND
classification	("classification" OR "classifier" OR "learning")

C. Study Selection

The selection process aims to identify the most relevant studies to address the RQs. After the search string portrayed in Table II, which had been applied to the digital libraries, numerous candidate papers were obtained, with one researcher retrieving candidate papers, and two others assessing them. The evaluation was based on titles, abstracts, keywords, and, in some instances, conclusions [21]. Moreover, the evaluation of the relevant papers utilized a predefined set of Inclusion Criteria (IC) and Exclusion Criteria (EC), both linked with the OR Boolean operator:

- IC1: Papers published in journals, conferences, and workshop proceedings.
- IC2: Only non-redundant papers, non-duplicate papers.
- IC3: Peer reviewed articles related to detailed multi-class imbalanced data issues.
- IC4: Relevant papers based on titles and abstracts and specifically focused on multi-class imbalanced databases.
- EC1: Non peer-reviewed papers not published in journals and proceedings.
- EC2: Publications considered as grey literature, i.e. working papers, presentations, and technical reports.
- EC3: Papers focusing on binary-class or discussing multiclass data but not imbalanced.

Highly relevant studies to multiclass imbalance data classification were successfully identified after the thorough inclusion and exclusion process, and were employed for subsequent data analysis.

D. Analysis

The acquired final set of multi-class imbalanced classification papers, was categorized according to the proposed categories of the techniques employed to address ambiguities, exhibited in Table III.

TABLE III. CATEGORIES OF TECHNIQUES

Techniques	Description
Oversampling	Mitigate imbalanced class distribution by replicating or creating examples of the minority class.
Undersampling	Create a subset of the original dataset by eliminating examples of the majority class.
Hybrid method	Combine the oversampling and undersampling methods.
Algorithm-level	Enhance the learning on minority classes by adapting specific classification algorithms.
Cost-sensitive	Impose greater penalties for misclassifying samples from the minority class as opposed to the majority class.
Ensemble-based	Create a more robust classifier by combining multiple base classifiers.
Data- and algorithm-level	Combine the data-level (oversampling or undersampling) and algorithm-level methods.

Subsequently, the details of each article were carefully reviewed to determine and categorize the specific problems addressed. The experimental section of each piece was analyzed to classify the type of dataset used in the context of multi-class imbalanced data classification.

E. Reporting

The final phase of SMS is referred to as the reporting process. Following the classification of the selected papers based on technique, problem, and dataset type during analysis, this study subsequently employed a bubble chart to visualize the outcomes, the problems solved, and the datasets utilized. In addition, the techniques followed within the 147 selected articles were detailed using the technique categorization proposed in this study. This offers valuable guidance for future research endeavors to address multi-class imbalanced data classification issues.

III. RESULT

A. Study Selection Results

The articles pertinent to the research on multi-class imbalance classification problems have been effectively identified by carefully implementing inclusion and exclusion operations. The refined outcomes, outlining the count of the retained articles for each database following the inclusion and exclusion processes, are demonstrated in Table IV.

TABLE IV. INCLUSION AND EXCLUSION RESULT

No.	Online Database	Selected	Relevant
1	ScienceDirect	4787	46
2	Scopus	775	35
3	IEEE Xplore	244	19
4	ACM Digital Library	825	4
5	Web of Science	533	22
TOTAL		7164	126

A total of 126 papers were identified in the selection stage. Subsequently, the snowball technique was applied to them, conducting a second search operation to distinguish any additional papers meeting the predefined criteria. Table V displays the total number of academic papers which were searched during this process, including the initial 126 articles and additional ones identified through the snowball technique. Table V offers an overview of the final number of articles analyzed in this study.

TABLE V. TOTAL ARTICLES

No.	Online database	Selected
1	Initially acquired papers	126
2	Snowball	21
TOTAL		147

B. Multi-class Imbalanced Data Classification Techniques (RQ1)

Figure 2 showcases the numerical distribution of the different technologies deployed in the selected 147 papers over the specified years, from 2018 to 2023. The algorithmic-level approach was most predominant among the selected studies, accounting for 30% of the cases. Ensemble-based and oversampling methods ranked second and third, constituting 25% and 23%, respectively. The hybrid, data-combined algorithm-level, and cost-sensitive methods followed closely. Notably, the undersampling method emerged as the least employed approach among the studies, with only two articles adopting it, representing a mere 1% of the overall research.

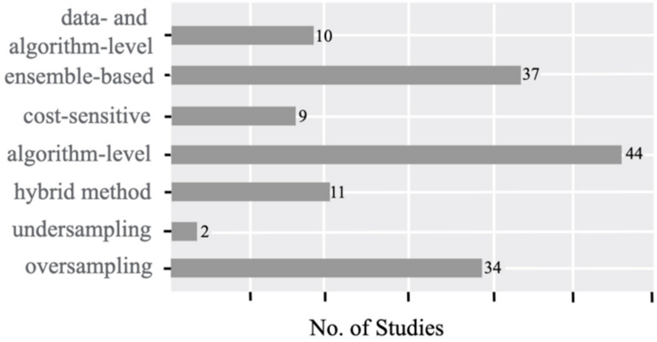

Fig. 2. Frequency of papers on multi-class imbalanced classification techniques (RQ1).

To gain a widespread understanding of the usage details and current trends of multi-class imbalanced data classification techniques, this study grouped those presented in Table III for the selected 147 studies into sub-types. Figure 3 illustrates the taxonomy of multi-class imbalanced data classification techniques, showing the distribution of algorithm-level methods with a more significant interest in Extreme Learning Machine (ELM) from the machine learning community since 12 out of 44 articles (27.3%) employ this approach. Furthermore, among the 12 selected research articles, 8 utilized weighted ELM (WELM) to address imbalanced datasets, representing the majority at 66.7% of the total. In addition, techniques focusing on Support Vector Machines (SVMs) are prevalent, with 7 out of the 44 (15.9%) selected research papers employing it. Nevertheless, out of the chosen studies, techniques other than Neural Networks (NN) and SVM, like Active Learning (AL), Broad Learning System (BLS), and tree-based method, were comparatively less popular. Specifically, three papers applied the technique AL, whereas only two papers employed the BLS and tree-based method techniques.

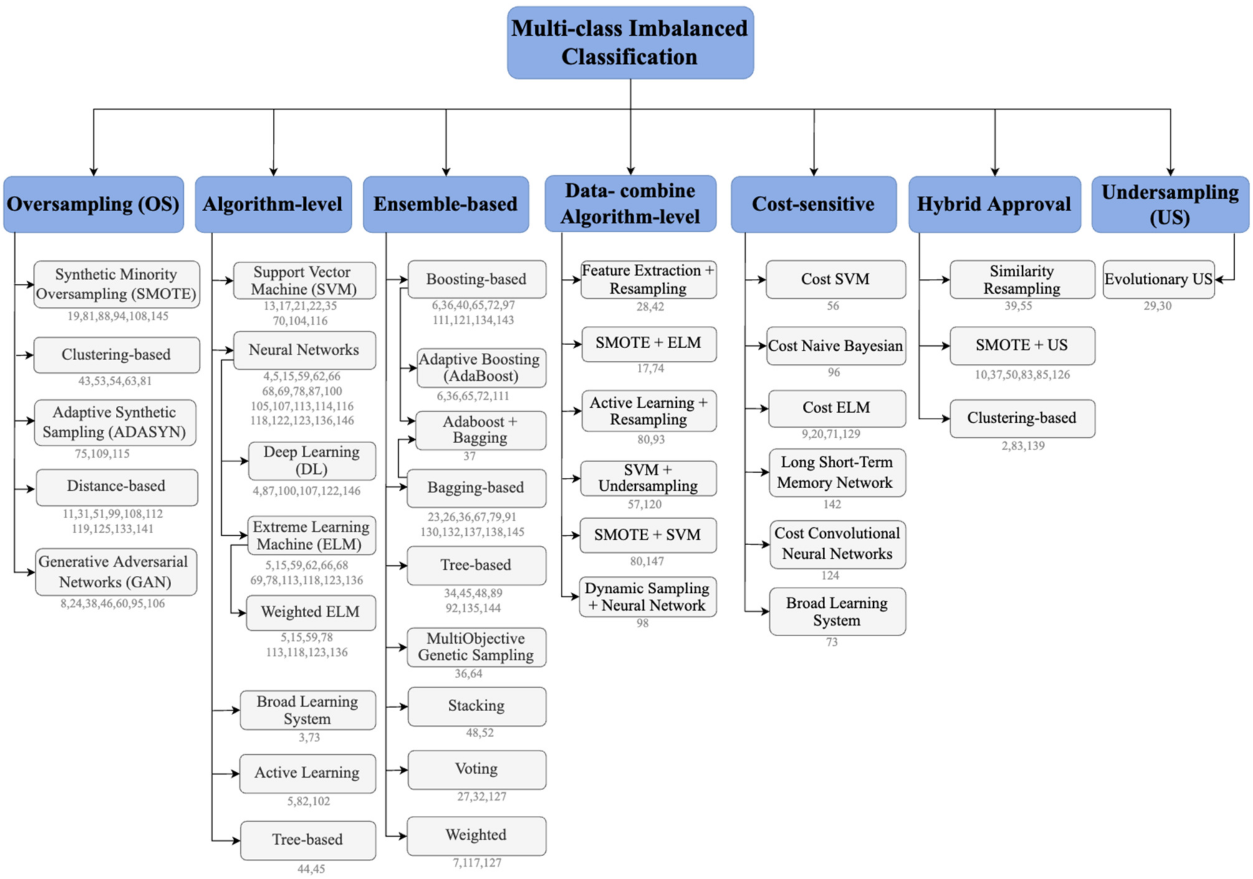

Fig. 3. Taxonomy of multi-class imbalanced data classification techniques.

There was an intensive use of bagging-based and boosting-based techniques for ensemble-based techniques, with 11 and 10 out of 37 studies identified, respectively, amounting to over then half of the total percentage (56.8%). The most frequently used boosting-based sub-technique was Adaboost, with 5 out of 10 studies classified as solution proposal papers. Interestingly, a survey combined bagging-based and boosting-based techniques, however, only one solution paper explored this approach. Among the remaining 16 studies, tree-based techniques indicate a preference over others, as 7 out of 16 papers (41.2%) employed this method.

For oversampling techniques, imbalanced data classification is predominantly accomplished through distance-based methods (29.4%), with 10 out of 34 articles utilizing this technique. Moreover, Generative Adversarial Networks (GANs), Synthetic Minority Oversampling (SMOTE), and clustering-based techniques were identified with a similar number of articles (7, 6, and 5, correspondingly), accounting for 20.6%, 17.6%, and 14.7% of the total. Furthermore, the remaining three papers employed the Adaptive Synthetic Sampling (ADASYN) method, making it the least widely used oversampling technique. The list of papers can be accessed at [22].

As depicted in Figure 3, in the remaining categories, the use of these four techniques is limited among the selected 147 studies, with 10 for combined algorithm-level techniques, 9 for cost-sensitive techniques, 11 for hybrid techniques, and 2 for undersampling techniques. Notably, SMOTE methods continue to feature prominently in these studies. Specifically, 6 out of 11 studies combined the SMOTE method with undersampling in the subcategory of hybrid forms. In contrast, in the subcategory of data combined algorithmic-level methods, 5 out of 10 studies combined other algorithm-level techniques to address imbalanced data classification. Furthermore, only 2 selected papers employed the undersampling technique for multiclass imbalanced data classification, indicating its low popularity.

C. Multi-class Imbalanced Data Classification Issues (RQ2)

There are numerous problems prevalent in multi-class imbalanced data classification. Still, this study presents only the relevant ones and those discussed in the papers without further clarifying categorization. Figure 4 illustrates the frequency of problems in multi-class imbalanced data classification.

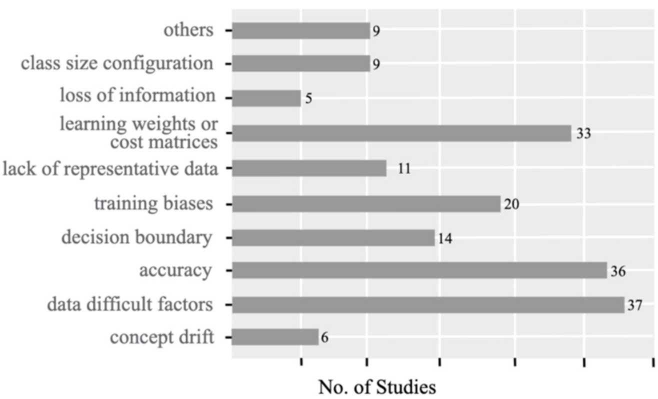

Fig. 4. Frequency of problems in multi-class imbalanced classification (RQ2).

The category designated as "others" refers to cases where no specific issues were explicitly mentioned but pertained to the challenges associated with multi-class imbalanced data classification. Some papers do not exclusively concentrate on solving a single problem but rather simultaneously address multiple related issues. Hence, the total number of articles addressing diverse problems exceeds the selected studies (147), indicating that some papers entail a range of research problems. As observed, 37 of the 147 articles (25.2%) predominantly focus on addressing the challenges posed by data difficulty factors, which renders it the most prevalent issue among the chosen studies. Issues related to accuracy and learning weights also frequently emerged, with 36 (24.5%) and 33 (22.4%) articles out of the selected ones, respectively.

Figure 5 illustrates the distribution of papers considering their percentage, focusing on diverse data difficulty factors within the subset of 37 papers. Half of the papers addressed class overlap issues, indicating the problem as one critical challenge in multi-class imbalanced data classification.

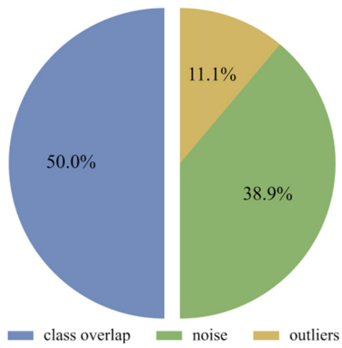

Fig. 5. Data difficulty factors.

D. Multi-class Imbalanced Data Classification Datasets (RQ3)

Figure 6 presents the classified datasets. Eleven categories emerge, and almost all of them use generalized datasets.

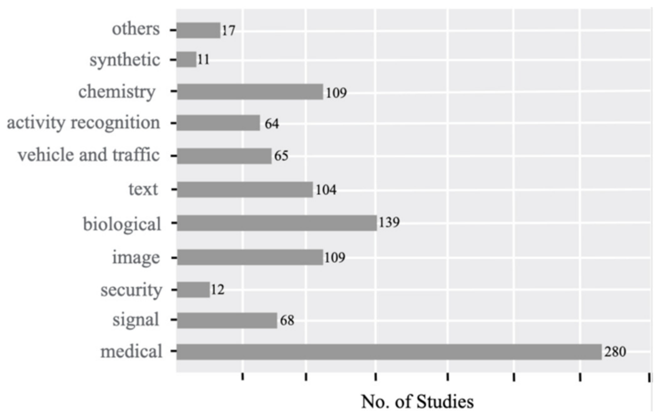

Fig. 6. Frequency of data types used in multi-class imbalanced classification (RQ3).

Specifically, a study may draw upon datasets spanning across multiple categories while incorporating various datasets within a single category. A comprehensive list of the datasets mentioned in the articles was compiled; for each, this study meticulously tallied the number of articles utilizing it. Next, these datasets were categorized into 11 distinct categories. Multiple datasets may be present within each category, as the article count was aggregated across these datasets to determine the cumulative dataset usage for each type. Based on the observations made, medical datasets were mainly used in the selected papers. Their dominance highlights the unique challenges in addressing the multi-class imbalanced data classification in the medical domain.

E. Mapping

To provide an overview of machine learning research in multi-class imbalanced data classification, a visual representation is offered in Figure 7. This mapping study illustrates the distribution of papers according to their employed techniques and addresses problems inherent to the types of datasets applied. Each paper is visually represented as a bubble, with its size and number reflecting the frequency of documents classified under each category. Since a paper may contribute to numerous techniques, problems, or dataset types, it is separated into various factors within each category. Consequently, the total number of paper counts in the bubble plot differs from the overall count of 147 relevant papers.

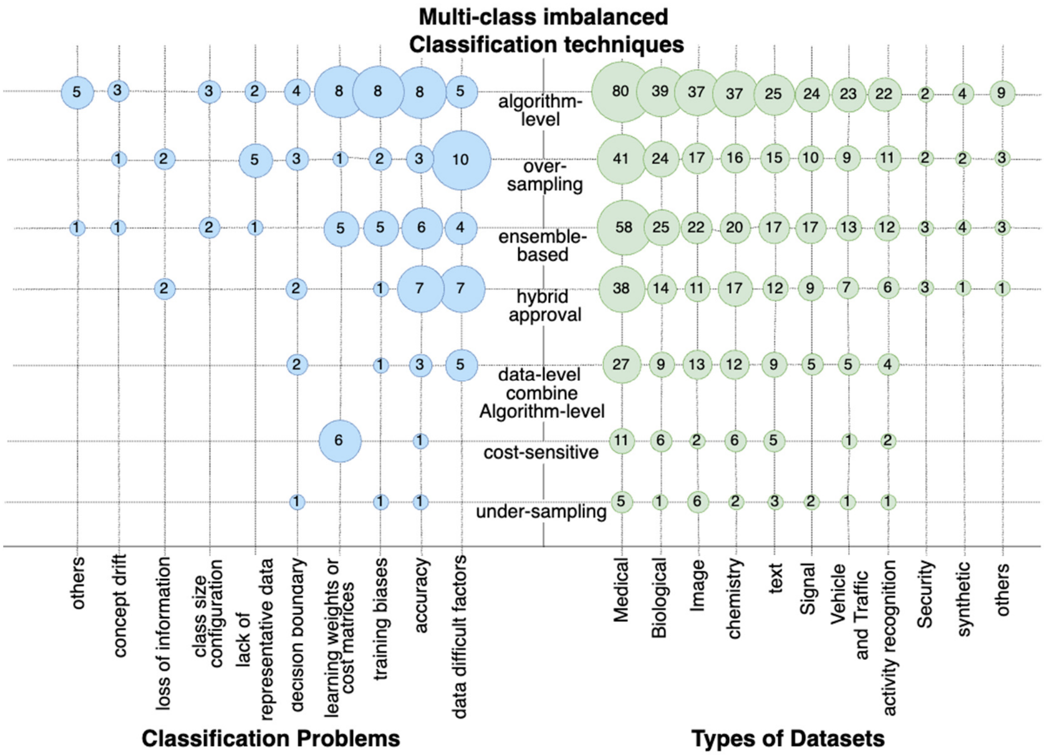

Fig. 7. Distribution of multi-class imbalanced data classification by techniques, problems, and dataset types.

As visible through the bubble plots, the algorithm-level techniques found to have been employed to address multi-class imbalanced data classification problems predominantly tackle three core challenges: accuracy, training bias, and learning weights or cost matrix. Still, when encountering challenges posed by data difficulty factors, the oversampling technique is the most widely adopted method, which is demonstrated by the fact that the majority of research papers was dedicated to addressing this issue. Moreover, regardless of the method employed to address the multi-class imbalanced data classification problem, it is clear that researchers consistently prioritize the exploration of medical datasets.

IV. DISCUSSION

This mapping study offers a comprehensive and much needed overview of multi-class imbalanced classification techniques, including the types of the datasets employed and the specific challenges. Researchers commonly prioritize following algorithm-level, ensemble-based, or oversampling methods as primary solutions for addressing the challenge of multi-class imbalanced classification (thereby answering RQ1). Out of these three, the algorithmic-level approach is found to be the most prominent one, with 44 of the 147 selected papers using it to address imbalanced data classification, representing the highest percentage (30%). The study's findings suggest that further exploring these methods may help alleviate challenges associated with multi-class imbalanced data classification.

Nevertheless, the results of this SMS in terms of the RQ2 indicate that the primary issues addressed differ when scholars utilize different methods. Algorithm-level and ensemble-based methods address issues related to accuracy, training biases, and learning weight, whereas oversampling techniques primarily focus on data difficulty factors. Moreover, this research has ascertained that data difficulty factors, accuracy, learning weights, and training bias are the most frequently mentioned multi-class imbalance issues, with 37, 36, 33, and 20 papers out of the total 147 attempting to solve it. Hence, researchers could achieve enhanced performance in multi-class imbalanced data classification by employing appropriate techniques to address these challenges.

Among data difficulty factors, the class overlap problem has received the greatest attention and has been extensively studied. Due to the prevalence of class overlap, which has been identified as the most common data complication addressed by oversampling methods, half of the 37 selected studies attempted to solve it. Additionally, the oversampling process is the most frequently adopted approach to address the issues of data difficulty factors, with 11 out of the 37 selected studies (30%) employing it. Therefore, class overlap is likely to be a core issue that needs to be faced when using oversampling methods to classify multi-class imbalanced data.

Furthermore, this study's mapping shows that most selected papers concentrate on medical datasets. Although a single paper may implement multiple datasets, the medical dataset type remains the most frequently used, with 280 occurrences in 147 articles, twice as many as the second most commonly utilized biological dataset (139), showcasing that the majority of the research efforts are towards solving multi-class imbalance classification problems in the medical domain (answering RQ3). Numerous researchers continue to study imbalanced medical dataset classification, probably because accurate disease diagnosis is critical for modernizing medical interventions. In future investigations, this paper suggests the potential of deploying advanced oversampling methods to simultaneously solve the multi-class imbalanced medical data classification and class overlap issue.

V. THREATS TO VALIDITY

Considering paper inclusion, it is inevitable that critical studies may be unintentionally omitted, and researcher bias may inadvertently influence the selection process. Thus, it is significant to address and discuss the main threats to the study's internal validity.

- Search string bias: Each digital database adheres to its distinct search protocol, encompassing various aspects like wildcard functionality, limitations on utilizing Boolean operators, and specific search parameters. Consequently, the search strings had to be meticulously tailored to ensure their adaptability to the unique specifications of each digital database, thereby optimizing and harmonizing the results across all the sources.
- Conclusion validity: In a systematic mapping study, various factors, including incorrect data extraction and potential omission of relevant studies, threaten conclusion validity. These factors can lead to conclusions that are erroneous or misleading. To mitigate this issue, bar and bubble plots as well as pie diagrams generated from the data were employed and comprehensively discussed. This approach effectively ensured traceability between the extracted data and the derived conclusions, ascertaining a transparent and verifiable representation of the findings.

VI. CONCLUSION

This paper conducted a systematic mapping study, examining the last five years (2018-2023) on the techniques, issues, and dataset usage employed in multi-class imbalanced data classification. After an extensive search and screening, 147 relevant studies that comprehensively analyzed the stated research questions were identified. The main contributions of this study include a taxonomy of techniques for multi-class imbalanced data classification and a thorough overview of multi-class imbalanced classification techniques, issues, and dataset types, which can give researchers a more comprehensive understanding of the field.

The ultimate findings of this systematic mapping study can be used to guide researchers in selecting appropriate techniques to address the challenge of multi-class imbalanced data classification. The results indicate that the algorithm level is the predominant technique employed in classifying multi-class imbalanced data, followed by ensemble-based and oversampling techniques. Meanwhile, based on the mapping results and data difficulty factors, class overlap emerged as the most extensively researched issue when attempting to group multi-class imbalanced data through oversampling techniques. The results of this mapping study reveal the widespread and prominent focus on classifying multi-class imbalanced data within the medical domain, considering that medical datasets are the most studied ones in the selected papers. According to our findings, there is a need for the research community to further explore and prioritize the development of solutions for handling multi-class imbalance in medical data, with specific emphasis on addressing the challenge of class overlap.

Given this potential future opportunity, this mapping study's findings are currently employed as the basis for developing a resampling framework dealing with multi-class imbalanced medical data. This framework will serve as a valuable resource for researchers seeking to address and mitigate the challenges of multi-class imbalanced data.

ACKNOWLEDGMENT

The authors would like to thank the Journal Support Fund, Universiti Teknologi MARA for their financial support to this study. We would also like to thank the College of Computing, Informatics & Mathematics, Universiti Teknologi MARA, Selangor, Malaysia for all the supports.

REFERENCES

[1] P. Branco, L. Torgo, and R. P. Ribeiro, "A Survey of Predictive Modeling on Imbalanced Domains," *ACM Computing Surveys*, vol. 49, no. 2, pp. 31:1-31:50, Aug. 2016, https://doi.org/10.1145/2907070.

[2] J. Forough and S. Momtazi, "Sequential credit card fraud detection: A joint deep neural network and probabilistic graphical model approach," *Expert Systems*, vol. 39, no. 1, 2022, Art. no. e12795, https://doi.org/10.1111/exsy.12795.

[3] A. Rezaeipanah and G. Ahmadi, "Breast Cancer Diagnosis Using Multi-Stage Weight Adjustment In The MLP Neural Network," *The Computer Journal*, vol. 65, no. 4, pp. 788–804, Apr. 2022, https://doi.org/10.1093/comjnl/bxaa109.

[4] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas, "Imbalance Problems in Object Detection: A Review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3388–3415, Oct. 2021, https://doi.org/10.1109/TPAMI.2020.2981890.

[5] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, https://doi.org/10.1109/TKDE.2008.239.

[6] D. Virmani, N. Jain, A. Srivastav, M. Mittal, and S. Mittal, "An Enhanced Binary Classifier Incorporating Weighted Scores," *Engineering, Technology & Applied Science Research*, vol. 8, no. 2, pp. 2853–2858, Apr. 2018, https://doi.org/10.48084/etasr.1962.

[7] R. M. Mathew and R. Gunasundari, "A Cluster-based Undersampling Technique for Multiclass Skewed Datasets," *Engineering, Technology & Applied Science Research*, vol. 13, no. 3, pp. 10785–10790, Jun. 2023, https://doi.org/10.48084/etasr.5844.

[8] L. Abdi and S. Hashemi, "To Combat Multi-Class Imbalanced Problems by Means of Over-Sampling Techniques," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 238–251, Jan. 2016, https://doi.org/10.1109/TKDE.2015.2458858.

[9] S. Wang and X. Yao, "Multiclass Imbalance Problems: Analysis and Potential Solutions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 1119–1130, Aug. 2012, https://doi.org/10.1109/TSMCB.2012.2187280.

[10] M. Liu, M. Dong, and C. Jing, "A modified real-value negative selection detector-based oversampling approach for multiclass imbalance problems," *Information Sciences*, vol. 556, pp. 160–176, May 2021, https://doi.org/10.1016/j.ins.2020.12.058.

[11] A. S. Palli, J. Jaafar, M. A. Hashmani, H. M. Gomes, and A. R. Gilal, "A Hybrid Sampling Approach for Imbalanced Binary and Multi-Class Data Using Clustering Analysis," *IEEE Access*, vol. 10, pp. 118639–118653, 2022, https://doi.org/10.1109/ACCESS.2022.3218463.

[12] B. Kitchenham, "Guidelines for performing Systematic Literature Reviews in Software Engineering," EBSE E-2007-01 Technical Report, Jan. 2007.

[13] A. Idri, H. Benhar, J. L. Fernández-Alemán, and I. Kadi, "A systematic map of medical data preprocessing in knowledge discovery," *Computer Methods and Programs in Biomedicine*, vol. 162, pp. 69–85, Aug. 2018, https://doi.org/10.1016/j.cmpb.2018.05.007.

[14] M. Hosni *et al.*, "A systematic mapping study for ensemble classification methods in cardiovascular disease," *Artificial Intelligence Review*, vol. 54, no. 4, pp. 2827–2861, Apr. 2021, https://doi.org/10.1007/s10462-020-09914-6.

[15] B. A. Tama and S. Lim, "Ensemble learning for intrusion detection systems: A systematic mapping study and cross-benchmark evaluation," *Computer Science Review*, vol. 39, Feb. 2021, Art. no. 100357, https://doi.org/10.1016/j.cosrev.2020.100357.

[16] J. Edward and M. M. Rosli, "A Systematic Mapping Study on Ensemble-Based Classifier," in *2021 IEEE International Conference on Computing (ICOCO)*, Nov. 2021, pp. 43–48, https://doi.org/10.1109/ICOCO53166.2021.9673563.

[17] S. Anwar, M. Mohd Rosli, and N. A. S. Abdullah, "Classification of Fault Prediction: A Mapping Study," *Pertanika Journal of Science and Technology*, vol. 30, pp. 2157–2171, May 2022, https://doi.org/10.47836/pjst.30.3.23.

[18] M. Maw, V. Balakrishnan, O. Rana, and S. D. Ravana, "Trends and Patterns of Text Classification Techniques: A Systematic Mapping Sudy," *Malaysian Journal of Computer Science*, vol. 33, no. 2, pp. 102–117, Apr. 2020, https://doi.org/10.22452/mjcs.vol33no2.2.

[19] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic Mapping Studies in Software Engineering," *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, vol. 17, Jun. 2008.

[20] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Information and Software Technology*, vol. 64, pp. 1–18, Aug. 2015, https://doi.org/10.1016/j.infsof.2015.03.007.

[21] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from applying the systematic literature review process within the software engineering domain," *Journal of Systems and Software*, vol. 80, no. 4, pp. 571–583, Apr. 2007, https://doi.org/10.1016/j.jss.2006.07.009.

[22] CodingLifeV, "CodingLifeV/MultiClassSMS." Mar. 05, 2024, [Online]. Available: https://github.com/CodingLifeV/MultiClassSMS.