Data Science 0 (0) 1
IOS Press

A Comprehensive Review of Evolutionary Sampling Techniques for Addressing Data Quality Problems in Imbalanced Data Classification

Fhira Nhita [a], Asniar [b], Isman Kurniawan [c] and Adiwijaya [d,*]
[a] *School of Computing, Telkom University, Bandung, West Java, Indonesia*
E-mail: fhiranhita@telkomuniversity.ac.id; ORCID: https://orcid.org/0000-0001-6050-8760
[b] *School of Applied Science, Telkom University, Bandung, West Java, Indonesia*
E-mail: asniar@telkomuniversity.ac.id; ORCID: https://orcid.org/0000-0003-3323-5466
[c] *School of Computing, Telkom University , Bandung, West Java, Indonesia*
E-mail: ismankrn@telkomuniversity.ac.id; ORCID: https://orcid.org/0000-0003-3485-3063
[d] *School of Computing, Telkom University, Bandung, West Java, Indonesia*
E-mail: adiwijaya@telkomuniversity.ac.id; ORCID: https://orcid.org/0000-0002-3518-7587

Abstract. With the rapid expansion of data, particularly in the form of data banks, numerous challenges have arisen, among which the issue of imbalanced data has become increasingly prominent. Generally, three main approaches are used to address imbalanced data, i.e., approaches at data-level, algorithm-level, and hybrid of both levels. The data-level approach, also known as sampling techniques, is widely adopted because the approach does not depend on specific classifier. Evolutionary computation has become a popular method in the sampling process, referred to as evolutionary sampling techniques, as has been effectively proven in various optimization tasks. Also, the imbalanced data issues are often related to data quality problems, such as noise and class overlapping. However, to the best of our knowledge, no survey has been performed that focused on evolutionary sampling techniques, particularly for handling noise and class overlapping problems. Hence, this paper presents a systematic literature review, offering a comprehensive discussion on evolutionary sampling techniques that focus on addressing noise and class overlapping problems. This survey identifies key challenges and opportunities, guiding future advancements in handling imbalanced data with evolutionary sampling techniques.

Keywords: Data banks, imbalanced data, evolutionary sampling techniques, data quality problems, systematic review

1. Introduction

1.1. Motivation

Imbalanced data classification remains a crucial problem because the performance of the classifier is frequently not satisfied due to a significant difference in sample sizes of classes[57, 58]. The problem is commonly encountered in a wide range of real-world applications, such as churn detection [12, 94],

*Corresponding author. E-mail: adiwijaya@telkomuniversity.ac.id.

financial fraud detection [84], medical diagnosis of cancer [80, 92], and internet or cyber attack detection [5, 95].

Generally, classification methods tend to minimize training errors for all samples, often leading to biased outcomes [53]. As a consequence, the minority classes are frequently misclassified due to their smaller size, resulting in significant losses [9, 100]. Addressing imbalanced data is challenging due to several data quality problems in these datasets, including noise, class overlapping, small disjunct, and dataset shift [34, 60, 63, 73]. Therefore, it is crucial for researchers to have a comprehensive understanding of imbalanced data characteristics, solution approaches, and future research opportunities.

In general, the approaches to address imbalanced data problems can be categorized into three levels, i.e., data-level, algorithm-level, and the hybrid of both levels [22, 41, 58, 81]. The data-level approach, also known as sampling or data preprocessing, has been widely utilized because the approach is independent of the used classifier [9]. This approach addresses imbalanced data through over sampling [14, 23], under sampling [32, 39, 51, 59, 64, 71, 95, 110], and hybrid sampling methods [10, 19, 60, 76, 91]. Even though these sampling methods have successfully addressed several imbalanced data problems, these methods still face numerous challenges in handling imbalanced data, particularly related to data quality problems, such as noise, class overlapping, small disjunct, or dataset shift [85].

To address those challenges, those sampling techniques are commonly combined with other methods, such as evolutionary computation (EC) algorithms [70]. The EC algorithms are known as optimization methods that are inspired by natural phenomena such as biological evolution and animal swarm behavior [6, 47]. Several EC algorithms are widely used in solving the optimization task, such as Genetic Algorithm (GA) [45, 108], Particle Swarm Optimization (PSO) [6, 27, 82, 102], and Differential Evolution (DE) [2, 46, 87]. By defining the sampling process as the optimization task, the implementation of EC in the sampling techniques, which is referred to as evolutionary sampling techniques, has become popular.

Regarding the necessity to address the imbalanced data problem, several survey papers reviewed various approaches, including sampling techniques [11, 23, 34–36, 50, 63, 83, 90, 96]. However, these studies do not cover evolutionary sampling techniques to address the imbalanced data classification, especially related to data quality problems. To date, only one survey paper by Pei et al. [70] discussed EC algorithms that give more focus on the algorithm-level approach. Moreover, Pei et al. [70] do not discuss how EC algorithms handle noise and class overlapping problems.

In this paper survey, we review the implementation of the evolutionary sampling techniques for addressing imbalanced data problems, especially those related to data quality problems, i.e., noise and class overlapping. We also discuss individual representations for evolutionary sampling techniques, open issues, and opportunities for future research.

1.2. Contributions

The main contributions of this survey paper are:

(1) Providing categorization of evolutionary sampling techniques for handling imbalanced data classification into three categories evolutionary sampling techniques for (1) handling general problems of imbalanced data, (2) addressing noise problems, and (3) addressing class overlapping problems.
(2) Discussing the individual representations that are being used in evolutionary sampling techniques for handling imbalanced data classification, both for general problems and specifically for handling noise and class overlapping problems.

(3) Discussing open issues and future research directions of evolutionary sampling techniques, including data quality problems, the design of individual representations, and research objectives such as training set selection, synthetic sample generation, and the optimization of sampling methods.

1.3. Paper organization

This survey paper is organized as follows: Section 2 presents the research methodology. Section 3 introduces the data quality problems in imbalanced data classification. Section 4 discusses the existing studies on evolutionary sampling techniques for addressing data quality problems in imbalanced data classification. Section 5 discusses individual representations in evolutionary sampling techniques, both for general problems and specifically for handling noise and class overlapping problems. Section 6 presents the taxonomy, open issues, and future research directions of evolutionary sampling techniques for handling imbalanced data classification, both for general problems and specifically for handling noise and class overlapping problems. Finally, Section 7 concludes the survey.

2. Research Methodology

This chapter describes the research methodology used in conducting a systematic literature review (SLR) that adheres to SLR guidelines derived from [101]. We conducted the SLR, which involves three steps: (i) formulating research questions (RQs), (ii) applying search strategy and selecting relevant literature, and (iii) extracting and analyzing the selected literature [4, 16, 21, 28, 65, 72, 89, 103, 106]. The process of SLR for this survey paper is shown in Fig. 1.

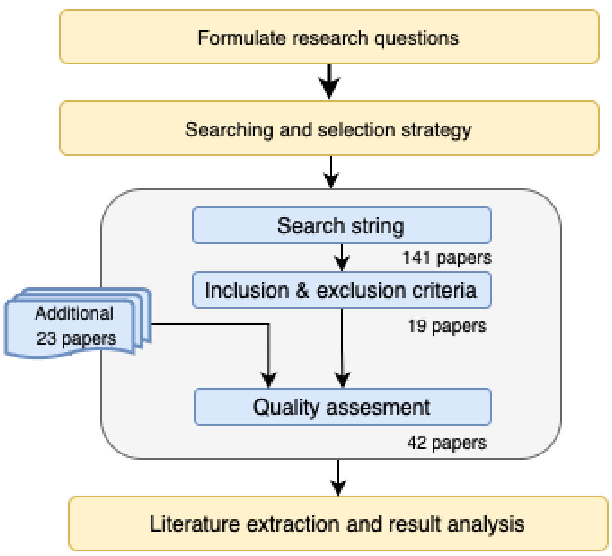

Fig. 1. Process of systematic literature review

2.1. Research Question

The first and most critical step in conducting a systematic literature review (SLR) is the formulation of research questions, as shown in Fig. 1. This step serves as the foundation of the SLR process, guiding it toward its objectives. Hence, we defined three research questions, as presented in Table 1.

Table 1
Research Questions (RQs)

RQs	Statements
RQ1	How are the evolutionary sampling techniques used to handle imbalanced data classification, both for general problems and specifically for handling noise and class overlapping problems?
RQ2	How are the individual representations designed for evolutionary sampling techniques to handle imbalanced data classification, both for general problems and specifically for handling noise and class overlapping problems?
RQ3	What are the open issues and future research directions in developing evolutionary sampling techniques to handle imbalanced data classification, both for general problems and specifically for handling noise and class overlapping problems?

2.2. Searching and Selection Strategy

The second step of the SLR process involves conducting a literature search and applying a selection strategy, as shown in Fig. 1. We performed a literature search in the Scopus repository using the following keywords: "(sampling OR balancing) AND (genetic OR evolutionary) AND imbalance AND (supervised OR classification)". We applied inclusion criteria to select papers for review as follows:

- Year: 2005 until 2023
- Language: English
- Accessibility: Documents available in scopus.com (user access: Telkom University, Date of search: 20 November 2023)
- Document type: PDF

We also applied the following exclusion criteria: (1) papers that were inaccessible through the institution's official account, and (2) papers whose content did not align with the use of evolutionary sampling techniques. Initially, we retrieved 141 papers, but only 19 met both the inclusion and exclusion criteria. The majority of the excluded papers were either inaccessible or involved the use of evolutionary sampling techniques in areas beyond the sampling process, such as feature selection, algorithm-level approaches, or as part of hybrid-level approaches.

Paul et al. [69] suggested using at least 40 papers as the minimum literature requirement for a systematic literature review. To meet the requirement, we manually added additional literature from Google Scholar and applied the inclusion and exclusion criteria. We obtained 23 additional papers, and thus the total number of papers reviewed in this survey is 42 papers. Then, the selected papers proceeded through a quality assessment step for the content evaluation of the full-text papers, including background problems, objectives, methodology, experimental results, and future works [72].

Finally, we presented the literature quality of the selected papers based on documents by year, types, and quartiles of articles as shown in Figs. 2, 3, and 4, respectively. Based on the distribution of documents by year, the majority of the selected papers were published between 2021 and 2023. In terms of document type, 71% of the selected papers were articles published in reputable international journals, with 80% of them appearing in Q1 journals and 17% in Q2 journals.

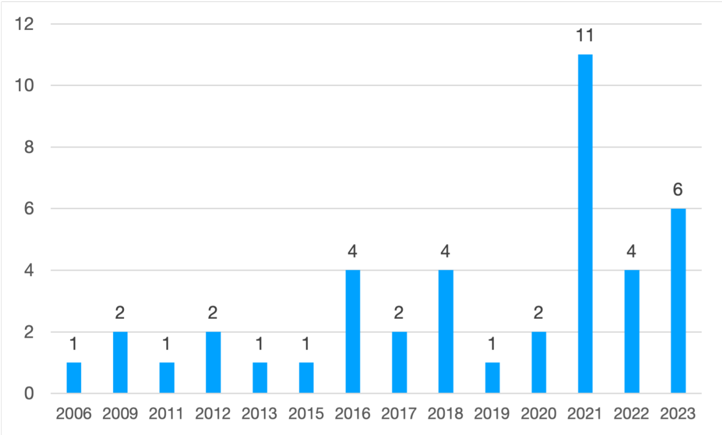

Fig. 2. Literature quality based on documents by year

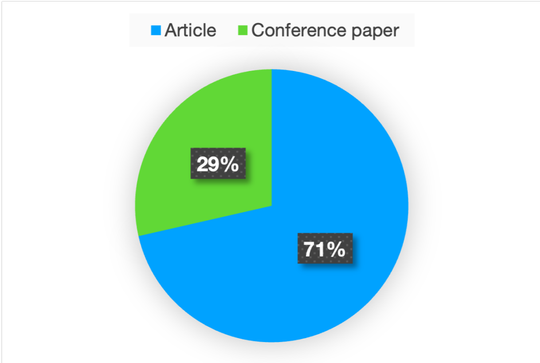

Fig. 3. Literature quality based on document by types

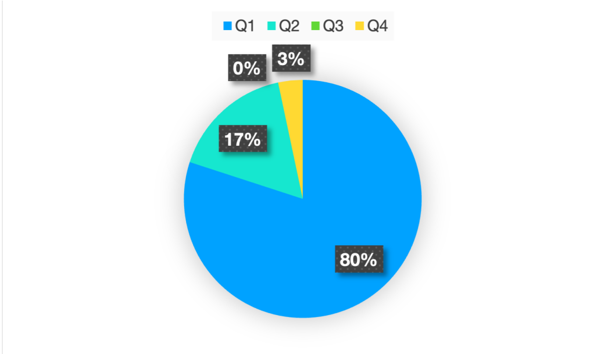

Fig. 4. Literature quality based on quartiles of articles

2.3. Literature Extract and Result Analysis

After collecting and selecting the literature, we conducted an information extraction and paper review to answer the three research questions (RQs), as shown in Table 1. The document of the review results are publicly available in the data repository: https://doi.org/10.5281/zenodo.14553584

3. Data quality problems in imbalanced data classification

Imbalanced data refers to a condition in a dataset where the number of a class sample significantly differs from the number of another class sample [53]. In the binary case, the class with a greater number of samples is called the majority class, while the class with a smaller number of samples is called the minority class [54]. The degree of imbalance condition in binary classification can be represented by the imbalance ratio (IR), as formulated in Eq. 1 [22, 53, 74],

$$IR_{binary\_class} = \frac{N_{\mathrm{maj}}}{N_{\mathrm{min}}} \tag{1}$$

where N_{maj} and N_{min} represent the number of samples in the majority and minority classes, respectively. A dataset is considered imbalanced if the IR value is greater than 1.5 [75].

The imbalanced ratio is a general problem in imbalanced data classification. However, several data quality problems also complicate classification tasks, such as noise, class overlapping, small disjunct, and dataset shift [34, 60, 63, 73].

3.1. Noise

The noise problem indicates the presence of noisy samples within the safe areas of other classes [77]. This disrupts classification performance, potentially reducing accuracy due to misclassification errors. An illustration of the noise problem is shown in Fig. 5.

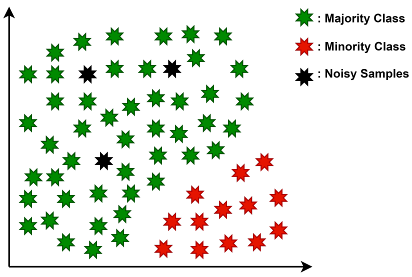

Fig. 5. Noise problem illustration

3.2. Class overlapping

Class overlapping occurs when samples from different classes mix in certain areas of the data representation, as illustrated in Fig. 6 [73]. This problem can be detected by using Fisher's discriminant ratio (F1) parameter. The smaller the F1 value, the greater the overlap condition in the dataset [63].

3.3. Small disjunct

The small disjunct condition occurs when minority samples are encapsulated within the majority class, as illustrated in Fig. 7 [77]. This condition poses a particular challenge that needs to be addressed by specific treatment [23].

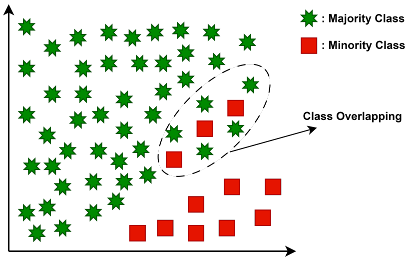

Fig. 6. Class overlapping illustration

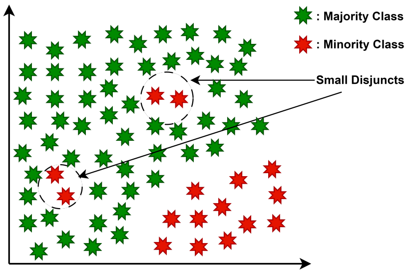

Fig. 7. Small disjunct illustration

3.4. Dataset shift

The dataset shift condition occurs when the distribution of data changes between the train and test sets. This condition leads to a decrease in classification performance on the test set. The illustration of the dataset shift condition is shown in Fig. 8 [34].

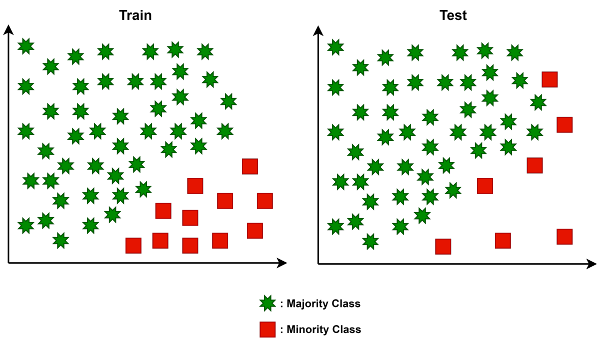

Fig. 8. Dataset shift illustration

4. Evolutionary sampling techniques for handling data quality problems in imbalanced data classification

This section answers the first research question by discussing the implementation of evolutionary sampling techniques for handling imbalanced data classification, both general problems and specific problems on noise and class overlapping issues.

4.1. Evolutionary sampling techniques for handling general imbalanced data problem

This first subsection discusses the reviewed studies that address general imbalanced data problems by focusing on balancing the sample sizes of majority and minority classes without addressing data quality problems, such as noise and class overlapping. In [29], proposed a wrapper-based random oversampling (WRO) method with GA to address imbalanced data issues. The wrapper-based approach has the advantage of utilizing feedback from the classifier during the over sampling process, ensuring that the generated synthetic data has been evaluated and improving classifier performance.

Sanguanmak and Hanskunatai [78] proposed the GA-DBSCAN-SMOTE (GADBSM) algorithm to address the limitations of the DBSM algorithm introduced in their previous study [79]. The results showed that incorporating GA into the optimization process significantly improved the performance of the DBSM algorithm. Ha and Lee [33] identified shortcomings in the under sampling process of the NearMiss and Similarity-based under sampling methods, prompting them to propose Genetic Algorithm-based Under Sampling (GAUS) for selecting informative majority samples.

Stanovov et al. [86] proposed a hybrid fuzzy evolutionary-based adaptive instance selection method, abbreviated as HEFCA, to select optimal train set samples. The goal of the HEFCA method is to adjust the sample selection probability so that the classification algorithm focuses on samples that are difficult to classify. In [44], GA was combined with SMOTE (GASMOTE) to address an issue in SMOTE related to uniform sampling rate for each sample in the minority class. The uniform sampling rate led to poor performance because each sample has a different contribution to the sampling process. Therefore, the GASMOTE method enhances the effectiveness of SMOTE by optimizing sampling rates for each sample in the minority class. Then, the optimized sampling rates were used to generate synthetic samples.

Jain et al. [42] proposed the Optimized Evolutionary Under Sampling (OEUS) method by optimizing the under sampling process using GA. They designed a new fitness function that assigns greater weight to sensitivity metrics than specificity. Ma et al. [62] proposed the Evolutionary Safe-level Synthetic Minority Over-sampling Technique (ESLSMOTE) that was derived from the Safe-level Synthetic Minority Over-sampling Technique (SLSMOTE), a SMOTE variant that generates synthetic samples in safe-level areas. ESLSMOTE generates synthetic samples in the areas that were dominated by nearest neighbors of minority samples. By using evolutionary computation, ESLSMOTE optimizes two main parameters of SLSMOTE consist of the number of nearest neighbors for the sampling process (k) and the number of nearest neighbors for calculating the safe-level process (c).

In [8], the GA algorithm was utilized for over sampling process to generate synthetic minority samples. The proposed method utilized Mahalanobis Distance (MD) to calculate the diversity measurement of minority samples, which differentiates it from other over sampling methods that primarily use Euclidean distance. Bui et al. [13] proposed a Cooperative Co-Evolutionary Software Defect Prediction (COESDP) framework. COESDP consists of three main stages, i.e., the balancing process using SMOTE-ENN hybrid sampling, sample selection optimization through a multi-population cooperative co-evolutionary approach (MPCA), and classification using ensemble learning. In the MPCA stage, GA is employed to drive the evolutionary process. The results showed that the multi-population approach achieved better classification performance compared to single-population methods in GA, PSO, and DE. However, the simultaneous evolution of multiple individuals adds complexity to the fitness value calculation, as it involves combining the best individuals from several sub-populations.

Jain et al. [43] proposed two methods consisting of GA-based undersampling and a multi-objective genetic algorithm (MOGA). The first method, GA-based undersampling, utilizes a fitness function that incorporates different weights for sensitivity and specificity, determined through a trial-and-error process. The second method, MOGA, enhances the performance of the first proposed method by optimizing the weight for both sensitivity and specificity. MOGA outperformed GA-based undersampling, indicating that MOGA can deliver superior classification performance for the minority class without sacrificing the performance of the majority class.

PSO algorithm is also widely used to address imbalanced data issues. In [102], a binary PSO (BPSO) was proposed to address imbalanced data issues in medical and biological datasets. This method focuses on selecting optimal samples from the majority samples, which are then concatenated with the minority samples to form a balanced dataset. García-López et al. [27] developed a PSO-based undersampling method that utilizes both wrapper and filter approaches. This study employed PSO to generate optimal balanced datasets, focusing on a comparative analysis of the wrapper and filter approaches as fitness functions.

Hu et al. [37] proposed an integrated SMOTE and PSO-based under sampling method to address the limitations of each method and to reduce the risk of variability. The proposed method applies SMOTE followed by PSO-based under sampling. SMOTE is used to over sampling the minority samples, while PSO is employed to under sampling the majority samples. The results showed that the hybrid sampling SMOTE and PSO performed significantly better compared to over sampling or under sampling approaches in most cases. Idris et al. [40] proposed an ensemble classification approach based on genetic programming dan AdaBoost where a fixed number of GP programs evolve per class in each iteration. To address the issue of imbalanced datasets, particle swarm optimization (PSO) based under sampling method is employed. This method selects discriminative samples from the majority class and combines them with samples from the minority class to create a balanced train set.

J. Li et al. [55] proposed three methods consisting of Swarm Dynamic Multi-Objective Rebalancing Algorithm (SDMORA), Swarm Instance Selection based on Swarm Dynamic Multi-Objective Rebalancing algorithm (SIsb-SDMORA), and Swarm Adaptive Clustered based Swarm Dynamic Multi-Objective Rebalancing algorithm (SaCb-SDMORA). The first method, SDMORA, utilized PSO algorithms to obtain optimal values for two SMOTE parameters, i.e., the over sampling rate (N) and the number of neighbors (K). SDMORA employs three objective functions i.e., accuracy, Kappa statistics, and balanced error rate. SDMORA focuses on the over sampling process, while the second method, SIsb-SDMORA, is integrated under sampling concepts with SDMORA, and the third method, SaCb-SDMORA, combines over sampling and under sampling.

Almomani et al. [6] proposed an evolutionary machine learning approach by utilizing binary particle swarm optimization (BPSO) to simultaneously optimize the parameters involved in the process, consisting of the features of datasets, parameters of the sampling methods, and parameters of the classification methods. The parameters of the sampling methods include the number of nearest neighbors and the sampling ratio. The study found that the best results were achieved using the SMOTE method, indicating that the optimizations performed by BPSO can be effective with SMOTE.

Shaw et al. [82] identified that over sampling and under sampling methods such as SMOTE and ENN have limitations, especially when dealing with datasets that have nominal features and are highly imbalanced. To address this challenge, they proposed a hybrid method, which combines Particle Swarm Optimization (PSO) and Ring Theory-based Evolutionary Algorithm (RTEA), abbreviated as RTPSO. The proposed method produced better performances on high-dimensional datasets compared to SMOTE because the synthetic samples are generated from variations of an original minority sample.

Differential Evolution (DE) is another optimization method frequently employed to address issues related to imbalanced data. DE has been utilized in several studies to tackle challenges associated with

SMOTE, as it is considered effective in overcoming SMOTE's limitations. In [2], DE is utilized to optimize SMOTE parameters. The proposed method, SMOTUNED, optimizes three key SMOTE parameters: the number of neighbors (k), the percentage of synthetic samples generated (m), and the distance function used (r). SMOTUNED explores various values for these parameters (k, m, r) across different datasets to maximize SMOTE's performance. The results emphasize that tuning SMOTE's parameters is crucial to identifying the best settings for each dataset.

The DE algorithm is also directly developed as an over sampling method, as demonstrated by [46], who designed a new oversampling method named Differential Evolution Based Oversampling approach for Highly Imbalanced Datasets (DEBOHID). This method employs the differential evolution algorithm to generate new candidate solutions in the oversampling process for highly imbalanced datasets. Furthermore, [49] implemented sixteen DE strategies for over sampling, with DEBOHID being one of them.

Another EC algorithm used to optimize parameters in over sampling methods is presented in [38], where an adaptive SMOTE was proposed by utilizing the state transition algorithm (STA) to optimize the best parameter pairs for SMOTE, such as the oversampling rate (N) and the number of nearest neighbors (k). Tao et al. [93] proposed the Evolutionary Synthetic Oversampling Technique (ESMOTE), which employs evolutionary strategies (ES) to optimize the over sampling ratio and the number of neighbors of SMOTE. These two parameters are crucial for generating synthetic minority samples. The over sampling ratio determines how many minority samples need to be synthesized to achieve balance, while the number of neighbors indicates how many nearest neighbors should be considered when generating synthetic minority samples.

García et al. [26] proposed Evolutionary Prototype Selection (EPS) to maximize accuracy and reduce the number of majority class samples. The EPS was performed by defining two scenarios based on two evolutionary methods i.e., Cross-generational elitist selection, Heterogeneous recombination, and Cataclysmic mutation (CHC) and Population-Based Incremental Learning (PBIL) for the under sampling process. This study designed a new fitness function based on the geometric mean and a penalization factor. The penalization factor is applied to the fitness function to ensure that the number of samples selected from each class remains balanced.

In a subsequent study, [25] designed the Evolutionary Undersampling (EUS) method to address imbalanced data problems. EUS employs evolutionary algorithms to select the majority samples. The study proposed eight experimental schemes according to objectives, selection schemes, and performance metrics. One of the proposed methods is EUS-CHC. The primary objective of this study was to evaluate the effectiveness of EUS methods in addressing the class imbalance problem. The concept of the EUS method has inspired other researchers to develop other EC methods in under sampling process such as [24, 52, 88].

On the other hand, Vluymans et al. [98] proposed an Evolutionary Prototype Reduction based Ensemble for Nearest Neighbor classification of Imbalanced Data (EPRENNID) to address the issues of imbalance and overfitting in the training process. The term prototype in this study represents samples in the general term. The proposed method focused on the under sampling process of the majority samples and allowed for the reduction of minority noise samples.

Table 2 summarizes several studies of evolutionary sampling techniques for handling general imbalanced data problems.

4.2. Evolutionary sampling techniques specifically addressing noise problem

This subsection discusses the reviewed studies that address imbalanced data problems while simultaneously giving special attention to noise issues. Abdouli et al. [1] proposed NCL+ to address the

Table 2
Overview of evolutionary sampling techniques for handling general imbalanced data problems

No	Year	Literature	Proposed method	Evolutionary algorithm(s)
1	2006	[26]	Evolutionary Prototype Selection	CHC, PBIL
2	2009	[102]	BPSO	PSO
3	2009	[25]	EUS	CHC, IGA
4	2012	[27]	PSO with wrapper and filter metric	PSO
5	2012	[29]	WRO	GA
6	2013	[24]	EUSBoost	CHC
7	2016	[33]	GAUS	GA
8	2016	[98]	EPRENNID	DE, SSMA
9	2016	[44]	GASMOTE	GA
10	2016	[78]	GADBSM	GA
11	2016	[86]	HEFCA IS	GA
12	2017	[42]	OEUS	GA
13	2018	[37]	SMOTE-PSO	PSO
14	2018	[18]	Autoencoder-PSO	PSO
15	2018	[2]	SMOTUNED	DE
16	2018	[88]	EUS-bag	GA
17	2019	[40]	GP-AdaBoost learning and PSO undersampling	PSO
18	2020	[38]	STA-SMOTE	STA
19	2021	[82]	RTPSO	PSO
20	2021	[52]	EUSC	CHC
21	2021	[93]	ESMOTE-CEM	ES
22	2021	[46]	DEBOHID	DE
23	2021	[49]	Boosting DE	DE
24	2021	[6]	SMOTE-tBPSO-SVM	PSO
25	2021	[62]	ESLSMOTE	GA
26	2021	[55]	Swarm Dynamic Multi-Objective Rebalancing Algorithm	PSO
27	2022	[8]	genetic algorithm-based oversampling	GA
28	2022	[13]	Cooperative Co-Evolutionary Software Defect Prediction	GA
29	2023	[43]	GA-based undersampling and multi-objective genetic algorithm	GA

limitation of the NCL algorithm by removing unwanted samples using CHC. Neighborhood cleaning rule (NCL) is one of the under sampling algorithms, proposed by Laurikkala [51]. Unlike NCL, which uses 3-NN in the under sampling process, NCL+ employs CHC for sample selection.

In [30], a Genetic Algorithm (GA) was used to identify a set of suspicious noise samples. After removing noise, adaptive sampling weights were calculated for each minority sample based on its proximity to the decision boundary, helping to address class overlapping. For these weights, samples closer to the decision boundary were assigned higher weights. The weights represented the probability of each sample being used for generating synthetic samples. Then, k-means clustering was applied to the minority class to create several clusters, within which synthetic samples were generated using SMOTE. The fitness function utilized was the geometric mean metric, calculated with k-NN and decision tree classifiers.

Junnan Li et al. [56] proposed the SMOTE-NaN-DE method, which combines SMOTE, Natural Neighbors, and Differential Evolution to balance data with less noise and borderline issues that commonly occur in SMOTE and its variants. The SMOTE-NaN-DE process begins with synthetic sample generation using SMOTE, followed by the detection of noise and borderline samples using Natural Neighbors. Finally, an optimization process with DE is conducted to maintain the balance of data ratios without removing samples detected as noise and borderline. However, DE improves these noise and borderline samples so that they can still contribute to classification and maintain data balance. This is the unique aspect of the SMOTE-NaN-DE method compared to others, which typically remove noise or borderline samples directly.

Z. Zhang and Li [107] proposed the Synthetic Minority Oversampling Technique based on Adaptive Local Mean Vectors and Improved Differential Evolution (SMOTE-LMVDE) to address the issue of overgeneralization that occurs when synthetic samples are noisy. SMOTE-LMVDE consists of three main stages: 1) noise detection using adaptive local mean vectors, 2) noise modification using differential evolution, and 3) an interpolation process to generate synthetic samples of the minority class. In the second stage, suspected noise are not removed but rather improved. The DE process involves a random difference between the suspicious noise and one of its nearest neighbors with the same class until the suspicious noise can be correctly classified by its nearest normal neighbor.

J. Zhang et al. [105] proposed SS_DEBOHID (Safe Set-based Differential Evolution on the Highly Imbalanced Dataset), which combines k-nearest neighbors (k-NN) and DE. This method addresses the weakness of the DEBOHID method in reducing noise during the synthetic minority sample generation process. DEBOHID generates synthetic samples across all minority samples, while SS_DEBOHID focuses on generating synthetic samples only within the safe area. The safe area is defined as the region dominated by minority samples, not within areas dominated by majority samples or the boundary regions between the two classes. This method employs two main stages: firstly, implementing k-NN to select minority samples in the safe area, where synthetic samples will be generated; and secondly, implementing DEBOHID to generate synthetic minority samples.

In [3], the Genetic-Novelty Oversampling Technique (GNOT) was proposed to generate synthetic minority samples. This method was introduced due to SMOTE's limitations in handling data with high imbalance ratios and outliers. The first step of GNOT is determining the outlier by using the Local Outlier Factor (LOF) to ensure that outlier samples are not selected in generating synthetic samples. The second step is generating synthetic minority samples according to the GA procedure. The crossover process in GNOT consists of two variants, i.e., GNOT (B) which uses barycentric crossover, and GNOT (L) which uses linear crossover. The use of barycentric crossover is an advantage of GNOT, as it ensures that the generated synthetic samples are located within the minority class region and in areas that are easily distinguishable from the majority class. This differs from the linear method, which may produce samples that are further from the original region, making it less effective in improving the representation of the minority class.

Wang et al. [99] proposed the natural neighbors-DEBOHID (NaN-DEBOHID) method to address the weakness of DEBOHID in handling noise. The main objective of this method is to identify better samples for synthetic sample generation and then remove noise from the new minority samples. The key difference between this method and DEBOHID is the process of identifying the natural neighbors of minority samples, which determines which samples will be used to generate synthetic samples (dense samples) and which are outliers to remove. The removal of outlier samples occurs after the synthetic samples are generated.

Z.-L. Zhang et al. [108] proposed an overproduce-and-choose synthetic example generation strategy based on evolutionary computation, called ESMOTE. The method comprises two stages to address SMOTE's issues, particularly the potential introduction of noise in the interpolation results. In the first stage, overproduction is performed using a modified SMOTE with a Gaussian distribution. During this stage, samples from both the minority and majority classes are selected to generate synthetic samples through SMOTE interpolation. In the second stage, the synthetic samples generated by SMOTE are selected using the CHC algorithm. The results demonstrate that ESMOTE's classification performance significantly outperforms SMOTE, SMOTE-Tomek Link, GA-SMOTE, and other comparative sampling methods.

Table 3 summarizes reviewed studies of evolutionary sampling techniques specifically addressing noise problems.

Table 3
Overview of evolutionary sampling techniques specifically addressing noise problem

No	Year	Literature	Proposed method	Evolutionary algorithm(s)
1	2015	[1]	NCL+	CHC
2	2021	[56]	SMOTE-NaN-DE	DE
3	2021	[30]	GA-SMOTE	GA
4	2022	[107]	SMOTE-LMVDE	DE
5	2023	[105]	SS_DEBOHID	DE
6	2023	[3]	GNOT	GA
7	2023	[99]	NaN-DEBOHID	DE
8	2023	[108]	ESMOTE	CHC

4.3. Evolutionary sampling techniques specifically addressing class overlapping problem

This section explores the reviewed studies that tackle imbalanced data issues while also focusing on class overlapping problems. Luengo et al. [61] focused on the implementation and analysis of data complexity metrics to evaluate the performance of sampling methods. This study found that only three measurements were the most informative i.e., F1 (maximum Fisher's discriminant ratio), N4 (non-linearity of the 1NN classifier), and L3 (non-linearity of the linear classifier by LP). This study implemented several sampling methods such as EUS-CHC [25] and SMOTE approach. The results indicated that EUS-CHC is more robust than the SMOTE approach in most cases. However, EUS-CHC has a notable drawback in terms of higher computational time complexity.

Zhu et al. [109] focused on removing the major samples in overlapping areas by eliminating samples in the overlapping region. Detection of overlapping samples is conducted by determining k-NN (k=1) for each majority sample. An evolutionary process using the CHC algorithm was performed on the samples of the majority class to decide whether the overlapping samples were to be selected or not. Then, the over sampling process was conducted using random over sampling (ROS) outside the evolutionary process. ROS was chosen to avoid generating new synthetic minority samples that may form new overlaps. The results showed that selecting ROS provided better classification performance compared to using no over sampling or SMOTE.

AlShourbaji et al. [7] proposed a novel method called HEOMGA, which combines the Heterogeneous Euclidean-Overlap Metric (HEOM) with a Genetic Algorithm (GA). This method aims to generate better synthetic samples from the minority class by using HEOM as the fitness function. HEOM measures the distance between minority samples and is considered more effective in handling diverse attributes, such as nominal and categorical data, compared to the commonly used euclidean distance metric. This method enables the selection of appropriate minority samples as input in the GA crossover process, resulting in more representative synthetic samples and reducing issues of overlapping and overfitting. This process is repeated until the number of minority samples in the current population is similar to the number of majority samples in the original dataset.

Gong et al. [31] proposed the Tomek link with genetic algorithm in the under sampling process (TEUS). This method was developed to address the issues of imbalance and overlapping by eliminating majority samples according to information contribution and overlap potential. TEUS consists of three steps, i.e., majority sample attribute estimation using the Tomek link, majority subgroup division, and under sampling process using genetic algorithm (GA).A subgroup is a set of majority samples that have similar values in terms of information contribution and overlap potential. Chromosome representation in GA uses an integer, where the value of each gene corresponds to a specific subgroup. TEUS applies GA to select samples from these subgroups and concatenate them with the minority samples.

Soltanzadeh et al. [85] addressed the issues of imbalance and class overlapping by conducting under sampling process using several EC algorithms to optimize the selection of majority samples. The proposed approach implemented three EC algorithms i.e., artificial bee colony (ABC), PSO, and GA. To evaluate the proposed approach, experiments were conducted on three types of datasets: synthetic datasets, real-world datasets, and large high-dimensional datasets. The results indicated that the ABC algorithm outperformed both PSO and GA. However, the experiments also showed that the proposed approach was not sensitive to the specific EC algorithm used.

One study by Junnan Li et al. [56] addresses both class overlapping and noise problems, as discussed in the previous subsection on noise problems. Therefore, they will not be discussed again in this subsection. The overview of evolutionary sampling techniques specifically addressing class overlapping problems in imbalanced data classification is shown in Table 4.

Table 4
Overview of evolutionary sampling techniques specifically handling class overlapping problem

No	Year	Literature	Proposed method	Evolutionary algorithm(s)
1	2011	[61]	EUSCHC	CHC
2	2020	[109]	EHSO	CHC
3	2021	[7]	HEOMGA	GA
4	2021	[56]	SMOTE-NaN-DE	DE
5	2022	[31]	TEUS	GA
6	2023	[85]	Meta-heuristic methods	ABC, PSO, GA

5. Individual representation of evolutionary sampling techniques

This section addresses the second research question concerning the individual representation of evolutionary sampling techniques, both for general problems and specifically for handling noise and class overlapping problems. Designing the individual representation is crucial to ensuring the effective operation of of evolutionary sampling techniques [25, 44, 62, 88].

The individual representation varies based on the used specific evolutionary algorithm and is related to the research objectives. Individual representation acts as a representation of candidate solutions, also known as solution representation in general, or chromosome representation in the context of genetic algorithms (GA) [8, 62, 88]. In this subsection, we elaborate the design of the individual representations of evolutionary sampling techniques into four types, i.e., binary, integer, real-valued, and combined representations.

Binary representation, where 0 indicates elimination and 1 indicates selection is frequently used in the reviewed studies, as designed in [85], [30], [18], [26], [25], [61], [24], [88], [52], [33], [42], [13], and [43]. Binary chromosomes were encoded with majority samples in overlapping regions, where 1 represents retained samples and 0 represents eliminated samples in [109]. Binary representation was represented by selected majority and minority samples in [98], while binary strings were used to represent majority samples as particles in [37]. Binary representation for two parameters of sampling methods designed in [62]. Binary chromosomes consist of minority samples used to generate new minority samples in [7]. Each gene in chromosomes encoded synthetic SMOTE minority samples for selection using binary representation in [108]. The SMOTE parameters to be optimized include the number of neighbors and the over sampling ratio in [6]. The number of neighbors of the SMOTE parameter is represented by four binary digits, corresponding to a decimal value between 0 and 15, reflecting the binary range from 0000 to 1111. Meanwhile, the over sampling ratio is represented by six binary digits, corresponding to a float value between 0.0 and 0.63, reflecting the binary range from 000000 to 111111.

In integer representation, the chromosome representation for fuzzy rules in [86] is represented as an integer string ranging from 0 to 14, which corresponds to fuzzy sets. Target vectors represented suspicious examples in integer value in [56]. Chromosomes were encoded with integers representing subgroups of majority samples for selection in [31]. Chromosomes represent integer values corresponding to the sampling rate of SMOTE for each minority sample in [44].

In real-valued representation, the chromosome representation consists of real numbers that indicate the minority samples selected for generating synthetic samples in [3]. Candidate solutions were represented by real-valued of three SMOTE parameters: k, m, and r in [2]. Studies on the DEBOHID method and its extensions have utilized real-valued vector representations for candidate solutions [46], [49], [105], [99]. Real-valued candidate solutions were represented as the area of minority samples for generating synthetic samples in [29]. The individual representation uses a real-valued type of two optimized SMOTE parameters in [38]. The individual representation is a real-valued decision vector involving multiple SMOTE parameters in [93]. The solution representation was represented with a real-valued vector for optimizing the two SMOTE parameters [55]. Each suspicious sample, identified as noise, is represented as a set of real-valued vectors in [107]. real-valued representation representing majority samples in [82], [102], [27], [40]. The individual representation is designed as samples of real-valued or integer representation in [8].

In combined representation, DBSM parameters were optimized using chromosomes composed of subchromosomes A, B, C, and D, with a total of 23 bits, including both binary and real values [78].

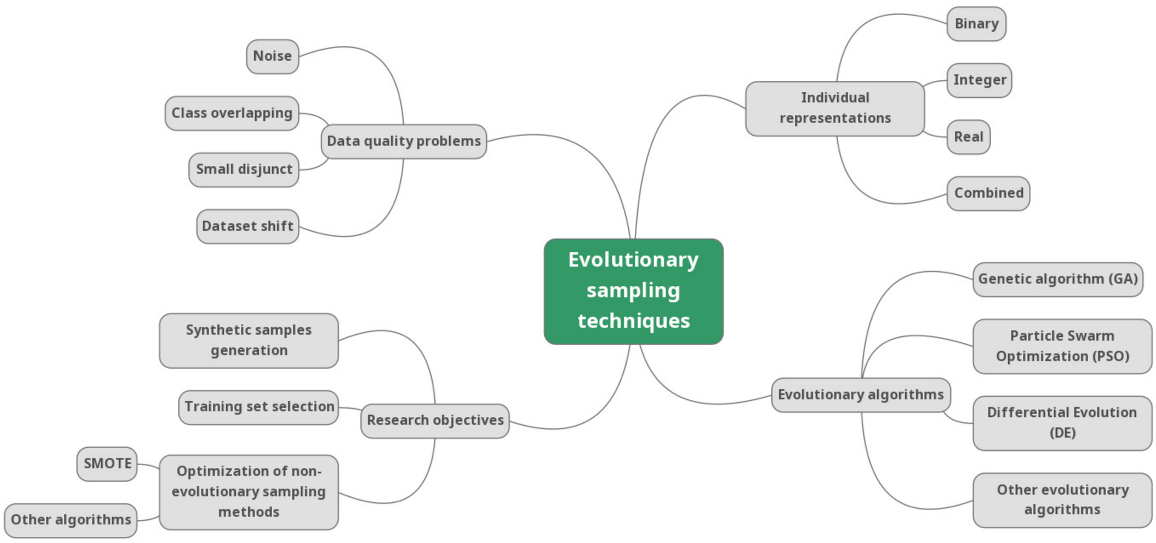

Fig. 9. Taxonomy of evolutionary sampling techniques

6. Open issues and future research direction

This section answers the third research question concerning the open issues of evolutionary sampling techniques and explores potential future research directions. We presented a taxonomy derived from the reviewed studies, into four categories i.e., data quality problems, individual representations, research objectives, and the evolutionary algorithms utilized, as illustrated in Fig. 9. Based on this taxonomy, several open issues and future research directions are discussed, including data quality problems, the design of individual representations, training set selection, synthetic sample generation, and the optimization of sampling methods.

6.1. Data quality problems

Addressing imbalanced data issues involves more than just balancing the number of samples between classes, as represented by the imbalance ratio. It also requires thorough consideration of the data quality problems that accompany the imbalanced data problem. Furthermore, our experimental study shows that the imbalance ratio does not sufficiently capture the complexity of the classification process [66, 67].

As shown in Tables 3 and 4, eight reviewed studies focus on addressing noise problems, while six studies focus on class overlapping problems. Only one study address both noise and class overlapping issues. These findings suggest that noise and class overlapping problems frequently accompany imbalanced data issues, requiring specific treatment. Conversely, no studies were found on evolutionary sampling techniques specifically addressing small disjunct and dataset shift, presenting an interesting challenge for future research.

6.2. Individual representation

Binary representation is the most commonly used for designing individual representations, possibly due to its ease of decoding from genotype to phenotype. This suggests an opportunity to explore design using alternative types of individual representations, while still aligning with the research objectives. In addition, based on Tables 2, 3, and 4, we identified three frequently used evolutionary algorithms i.e., GA, PSO, and DE, applied in 18, 9, and 9 studies, respectively. This indicates the effectiveness of GA in evolutionary sampling techniques for handling imbalanced data, particularly addressing noise and class overlapping issues.

6.3. Training set selection and synthetic samples generation

The reviewed studies of evolutionary sampling techniques employed various research objectives, including training set selection and synthetic sample generation. There are twenty-six studies focused on training set selection, while only eight studies focus on synthetic sample generation. This highlights open research opportunities for generating synthetic samples directly using evolutionary sampling techniques.

The process of synthetic sample generation is still predominantly conducted using SMOTE, which is then optimized by evolutionary algorithms. Furthermore, optimizing synthetic samples generated by methods other than SMOTE is also a potential area for exploration. In addition, hybrid techniques present opportunities for future research by combining evolutionary sampling techniques with other non-evolutionary sampling methods.

6.4. Optimization of sampling methods

Evolutionary sampling techniques play a crucial role in optimizing parameters in sampling methods. Among the reviewed studies, six studies focus on optimizing SMOTE parameters. The commonly optimized SMOTE parameters include the sampling ratio for each sample, the number of neighbors, and the percentage of synthetic samples generated.

On the other hand, there are two studies that focus on optimizing the other sampling methods. The first study optimized the SLSMOTE method by adjusting the number of nearest neighbors for both the sampling process and the safe-level calculation [62]. The second study optimized the DBSCAN method for two parameters i.e., the cluster radius and the minimum number of neighbors within a cluster [78]. This presents an opportunity for future research to develop evolutionary sampling techniques for optimizing sampling methods other than SMOTE.

7. Conclusions

This survey paper provides a comprehensive review of evolutionary sampling techniques for handling imbalanced data problems, specifically addressing noise and class overlapping issues. This is crucial because most real-world datasets face these challenges. We presented an overview of several studies that implement evolutionary sampling techniques both for general problems and specifically for handling noise and class overlapping problems. This survey revealed that most studies focus on evolutionary sampling techniques for handling noise and class overlapping problems, while studies on evolutionary sampling techniques for addressing small disjunct and dataset shift remain underexplored.

We have also discussed aspects of individual representations in evolutionary sampling techniques. Most studies have utilized binary representation for individual representations, which is also related to GA, the most commonly used evolutionary sampling technique. Based on research objectives, the implementation of evolutionary sampling techniques to generate synthetic samples and optimize parameters for sampling methods other than SMOTE remains rare. Finally, we present a taxonomy and discuss open issues and future research directions for evolutionary sampling techniques, individual representations, training set selection and synthetic sample generation, and optimization of sampling methods.

Acknowledgment

The authors would like to thank Direktorat Jenderal Pendidikan tinggi, Riset, dan Teknologi for financial support in this research (grant number: 180/E5/PG.02.00.PL/2023), and also for Telkom University.

References

[1] N.O.A. Abdouli, Z. Aung, W.L. Woon and D. Svetinovic, Tackling Class Imbalance Problem in Binary Classification using Augmented Neighborhood Cleaning Algorithm, in: *Information Science and Applications*, Springer Berlin Heidelberg, 2015, pp. 827–834.

[2] A. Agrawal and T. Menzies, Is "better data" better than "better data miners"?, in: *Proceedings of the 40th International Conference on Software Engineering*, ACM, New York, NY, USA, 2018.

[3] H. Ait Addi, R. Ezzahir and N. Boukhlik, Genetic-Novelty Oversampling Technique for Imbalanced Data, in: *Proceedings of the 6th International Conference on Big Data and Internet of Things*, Springer International Publishing, 2023, pp. 171–185.

[4] S. Al Faraby, A. Adiwijaya and A. Romadhony, Review on Neural Question Generation for Education Purposes, *International Journal of Artificial Intelligence in Education* (2023).

[5] B. Alabduallah, M. Maray, N. Alruwais, R. Alabdan, A.A. Darem, F.S. Alallah, R. Alsini and A. Yafoz, Class imbalanced data handling with cyberattack classification using Hybrid Salp Swarm Algorithm with deep learning approach, *Alex. Eng. J.* **106** (2024), 654–663.

[6] I. Almomani, R. Qaddoura, M. Habib, S. Alsoghyer, A. Al Khayer, I. Aljarah and H. Faris, Android Ransomware Detection Based on a Hybrid Evolutionary Approach in the Context of Highly Imbalanced Data, *IEEE Access* **9** (2021), 57674–57691.

[7] I. AlShourbaji, N. Helian, Y. Sun and M. Alhameed, Anovel HEOMGA Approach for Class Imbalance Problem in the Application of Customer Churn Prediction, *SN Computer Science* **2**(6) (2021), 464.

[8] C. Arun and C. Lakshmi, Genetic algorithm-based oversampling approach to prune the class imbalance issue in software defect prediction, *Soft Comput.* **26**(23) (2022), 12915–12931.

[9] Asniar, N.U. Maulidevi and K. Surendro, SMOTE-LOF for noise identification in imbalanced data classification, *Journal of King Saud University - Computer and Information Sciences* **34**(6, Part B) (2022), 3413–3423.

[10] G.E.A.P.A. Batista, R.C. Prati and M.C. Monard, A study of the behavior of several methods for balancing machine learning training data, *SIGKDD Explor. Newsl.* **6**(1) (2004), 20–29.

[11] P. Branco, L. Torgo and R.P. Ribeiro, A survey of predictive modeling on imbalanced domains, *ACM Computing Surveys* **49**(2) (2016).

[12] A. Bugajev, R. Kriauziene, O. Vasilecas and V. Chadyšas, The Impact of Churn Labelling Rules on Churn Prediction in Telecommunications, *Informatica* **33**(2) (2022), 247–277.

[13] L.T. Bui, V. Van Truong, B. Van Pham and V.A. Phan, A multi-population coevolutionary approach for Software defect prediction with imbalanced data, in: *2022 14th International Conference on Knowledge and Systems Engineering (KSE)*, IEEE, 2022, pp. 1–6.

[14] N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, *jair* **16** (2002), 321–357.

[15] V.K. Chennuru and S.R. Timmappareddy, Simulated annealing based undersampling (SAUS): a hybrid multi-objective optimization method to tackle class imbalance, *Appl. Intell.* **52**(2) (2022), 2092–2110.

[16] A. Cherif, A. Badhib, H. Ammar, S. Alshehri, M. Kalkatawi and A. Imine, Credit card fraud detection in the era of disruptive technologies: A systematic review, *J. King Saud Univ. - Comput. Inf. Sci.* **35**(1) (2023), 145–174.

[17] T.K. Dang, T.C. Tran, L.M. Tuan and M.V. Tiep, Machine learning based on resampling approaches and deep reinforcement learning for credit card fraud detection systems, *Applied Sciences (Switzerland)* **11**(21) (2021).

[18] M. Daoud and M. Mayo, A Novel Synthetic Over-Sampling Technique for Imbalanced Classification of Gene Expressions Using Autoencoders and Swarm Optimization, in: *AI 2018: Advances in Artificial Intelligence*, Springer International Publishing, 2018, pp. 603–615.

[19] D. Datta, P.K. Mallick, J. Shafi, J. Choi and M.F. Ijaz, Computational Intelligence for Observation and Monitoring: A Case Study of Imbalanced Hyperspectral Image Data Classification, *Comput. Intell. Neurosci.* **2022** (2022).

[20] R. Eberhart and J. Kennedy, A new optimizer using particle swarm theory, in: *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, IEEE, 1995, pp. 39–43.
[21] Z. Ellaky, F. Benabbou and S. Ouahabi, Systematic literature review of social media bots detection systems, *J. King Saud Univ. - Comput. Inf. Sci.* **35**(5) (2023), 101551.
[22] M. Fattahi, M.H. Moattar and Y. Forghani, Improved cost-sensitive representation of data for solving the imbalanced big data classification problem, *Journal of Big Data* **9**(1) (2022), 1–24.
[23] A. Fernández, S. García, F. Herrera and N.V. Chawla, SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary, *J. Artif. Intell. Res.* **61** (2018), 863–905.
[24] M. Galar, A. Fernández, E. Barrenechea and F. Herrera, EUSBoost: Enhancing ensembles for highly imbalanced datasets by evolutionary undersampling, *Pattern Recognit.* **46**(12) (2013), 3460–3471.
[25] S. García and F. Herrera, Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy, *Evol. Comput.* **17**(3) (2009), 275–306.
[26] S. García, J.R. Cano, A. Fernández and F. Herrera, A Proposal of Evolutionary Prototype Selection for Class Imbalance Problems, in: *Intelligent Data Engineering and Automated Learning – IDEAL 2006*, Springer Berlin Heidelberg, 2006, pp. 1415–1423.
[27] S. García-López, J.A. Jaramillo-Garzón, J.C. Higuita-Vásquez and C.G. Castellanos-Domínguez, Wrapper and filter metrics for pso-based class balance applied to protein subcellular localization, in: *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms*, SciTePress - Science and and Technology Publications, 2012.
[28] N. Ghaniaviyanto Ramadhan, Adiwijaya, W. Maharani and A. Akbar Gozali, Chronic diseases prediction using machine learning with data preprocessing handling: A critical review, *IEEE Access* **12** (2024), 80698–80730.
[29] A. Ghazikhani, H.S. Yazdi and R. Monsefi, Class imbalance handling using wrapper-based random oversampling, in: *20th Iranian Conference on Electrical Engineering (ICEE2012)*, IEEE, 2012, pp. 611–616.
[30] J. Gong, A Novel Oversampling Technique for Imbalanced Learning Based on SMOTE and Genetic Algorithm, in: *Neural Information Processing*, Springer International Publishing, 2021, pp. 201–212.
[31] P. Gong, J. Gao and L. Wang, A hybrid evolutionary under-sampling method for handling the class imbalance problem with overlap in credit classification, *J. Syst. Sci. Syst. Eng.* **31**(6) (2022), 728–752.
[32] S. Gupta, L. Goel, A. Singh, A. Prasad and M.A. Ullah, Psychological Analysis for Depression Detection from Social Networking Sites, *Comput. Intell. Neurosci.* **2022** (2022).
[33] J. Ha and J.-S. Lee, A New Under-Sampling Method Using Genetic Algorithm for Imbalanced Data Classification, in: *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, IMCOM '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 1–6.
[34] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue and G. Bing, Learning from class-imbalanced data: Review of methods and applications, *Expert Syst. Appl.* **73** (2017), 220–239.
[35] M.H.A. Hamid, M. Yusoff and A. Mohamed, Survey on Highly Imbalanced Multi-class Data, *International Journal of Advanced Computer Science and Applications* **13**(6) (2022), 211–229.
[36] K.M. Hasib, M.S. Iqbal, F.M. Shah, J. Al Mahmud, M.H. Popel, M.I.H. Showrov, S. Ahmed and O. Rahman, A Survey of Methods for Managing the Classification and Solution of Data Imbalance Problem (2020).
[37] Z. Hu, R. Chiong, I. Pranata, Y. Bao and Y. Lin, Malicious web domain identification using online credibility and performance data by considering the class imbalance issue, *Industrial Management & Data Systems* **119**(3) (2018), 676–696.
[38] Z. Huang, C. Yang, X. Chen, K. Huang and Y. Xie, Adaptive over-sampling method for classification with application to imbalanced datasets in aluminum electrolysis, *Neural Comput. Appl.* **32**(11) (2020), 7183–7199.
[39] A. Huč, J. Šalej and M. Trebar, Analysis of machine learning algorithms for anomaly detection on edge devices, *Sensors* **21**(14) (2021).
[40] A. Idris, A. Iftikhar and Z.u. Rehman, Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling, *Cluster Comput.* **22**(S3) (2019), 7241–7255.
[41] A. Jadhav, S.M. Mostafa, H. Elmannai and F.K. Karim, An Empirical Assessment of Performance of Data Balancing Techniques in Classification Task, *Applied Sciences (Switzerland)* **12**(8) (2022).
[42] A. Jain, S. Ratnoo and D. Kumar, Addressing class imbalance problem in medical diagnosis: A genetic algorithm approach, in: *2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC)*, IEEE, 2017, pp. 1–8.
[43] A. Jain, S. Ratnoo and D. Kumar, A novel multi-objective genetic algorithm approach to address class imbalance for disease diagnosis, *Int. J. Inform. Technol.* **15**(2) (2023), 1151–1166.
[44] K. Jiang, J. Lu and K. Xia, A novel algorithm for imbalance data classification based on genetic algorithm improved SMOTE, *Arab. J. Sci. Eng.* **41**(8) (2016), 3255–3266.
[45] P. Kaur and A. Gosain, FF-SMOTE: A Metaheuristic Approach to Combat Class Imbalance in Binary Classification, *Appl. Artif. Intell.* **33**(5) (2019), 420–439.

[46] E. Kaya, S. Korkmaz, M.A. Sahman and A.C. Cinar, DEBOHID: A differential evolution based oversampling approach for highly imbalanced datasets, *Expert Syst. Appl.* **169** (2021), 114482.
[47] J. Kennedy and R. Eberhart, Particle swarm optimization, in: *Proceedings of ICNN'95 - International Conference on Neural Networks*, Vol. 4, IEEE, 1995, pp. 1942–1948 vol.4.
[48] J. Kennedy and R.C. Eberhart, A discrete binary version of the particle swarm algorithm, in: *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, Vol. 5, IEEE, 1997, pp. 4104–4108 vol.5.
[49] S. Korkmaz, M.A. Şahman, A.C. Cinar and E. Kaya, Boosting the oversampling methods based on differential evolution strategies for imbalanced learning, *Appl. Soft Comput.* **112** (2021), 107787.
[50] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, *Progress in Artificial Intelligence* **5**(4) (2016), 221–232.
[51] J. Laurikkala, Improving Identification of Difficult Small Classes by Balancing Class Distribution, in: *Artificial Intelligence in Medicine*, Springer Berlin Heidelberg, 2001, pp. 63–66.
[52] H.L. Le, D. Landa-Silva, M. Galar, S. Garcia and I. Triguero, EUSC: A clustering-based surrogate model to accelerate evolutionary undersampling in imbalanced classification, *Appl. Soft Comput.* **101** (2021), 107033.
[53] J. Lee, D. Jung, J. Moon and S. Rho, Advanced R-GAN: Generating anomaly data for improved detection in imbalanced datasets using regularized generative adversarial networks, *Alex. Eng. J.* **111** (2025), 491–510.
[54] D.-C. Li, S.-C. Chen, Y.-S. Lin and W.-Y. Hsu, A Novel Classification Method Based on a Two-Phase Technique for Learning Imbalanced Text Data, *Symmetry* **14**(3) (2022).
[55] J. Li, Y. Wu, S. Fong, R.K. Wong, V.W. Chu, K.-L. Ong and K.K.L. Wong, Dynamic swarm class rebalancing for the process mining of rare events, *J. Supercomput.* **77**(7) (2021), 7549–7583.
[56] J. Li, Q. Zhu, Q. Wu, Z. Zhang, Y. Gong, Z. He and F. Zhu, SMOTE-NaN-DE: Addressing the noisy and borderline examples problem in imbalanced classification by natural neighbors and differential evolution, *Knowledge-Based Systems* **223** (2021), 107056.
[57] L. Li, K. Zhao, R. Sun, J. Gan, G. Yuan and T. Liu, Parameter-Free Extreme Learning Machine for Imbalanced Classification, *Neural Process. Letters* **52**(3) (2020), 1927–1944.
[58] C. Liao and M. Dong, ACWGAN: AN AUXILIARY CLASSIFIER WASSERSTEIN GAN-BASED OVERSAMPLING APPROACH FOR MULTI-CLASS IMBALANCED LEARNING, *Journal of Innovative Computing, Information and Control* (2022).
[59] A. Lombardi, N. Amoroso, L. Bellantuono, S. Bove, M.C. Comes, A. Fanizzi, D. La Forgia, V. Lorusso, A. Monaco, S. Tangaro, R. Bellotti and R. Massafra, Accurate Evaluation of Feature Contributions for Sentinel Lymph Node Status Classification in Breast Cancer, *Applied Sciences (Switzerland)* **12**(14) (2022).
[60] V. López, A. Fernández, J.G. Moreno-Torres and F. Herrera, Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics, *Expert Syst. Appl.* **39**(7) (2012), 6585–6608.
[61] J. Luengo, A. Fernández, S. García and F. Herrera, Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling, *Soft Computing* **15**(10) (2011), 1909–1936.
[62] J. Ma, D.O. Afolabi, J. Ren and A. Zhen, Predicting Seminal Quality via Imbalanced Learning with Evolutionary Safe-Level Synthetic Minority Over-Sampling Technique, *Cognit. Comput.* **13**(4) (2021), 833–844.
[63] S. Maheshwari, R.C. Jain and R.S. Jadon, An insight into rare class problem: Analysis and potential solutions, *Journal of Computer Science* **14**(6) (2018), 777–792.
[64] G. Mutanov, V. Karyukin and Z. Mamykova, Multi-class sentiment analysis of social media data with machine learning algorithms, *Computers, Materials and Continua* **69**(1) (2021), 913–930.
[65] S. Nazah, S. Huda, J. Abawajy and M.M. Hassan, Evolution of dark web threat analysis and detection: A systematic approach, *IEEE Access* **8** (2020), 171796–171819.
[66] F. Nhita, Adiwijaya and I. Kurniawan, Improvement of Imbalanced Data Handling: A Hybrid Sampling Approach by using Adaptive Synthetic Sampling and Tomek links, in: *2023 Eighth International Conference on Informatics and Computing (ICIC)*, 2023, pp. 1–5. doi:10.1109/ICIC60109.2023.10381929.
[67] F. Nhita, Adiwijaya and I. Kurniawan, Performance and Statistical Evaluation of Three Sampling Approaches in Handling Binary Imbalanced Data Sets, in: *2023 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, 2023, pp. 420–425.
[68] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J.M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting and D. Moher, The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, *Rev. Esp. Cardiol.* **74**(9) (2021), 790–799.
[69] J. Paul, W.M. Lim, A. O'Cass, A.W. Hao and S. Bresciani, Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR), *Int. J. Consum. Stud.* **45**(4) (2021).

[70] W. Pei, B. Xue, M. Zhang, L. Shang, X. Yao and Q. Zhang, A Survey on Unbalanced Classification: How Can Evolutionary Computation Help?, *IEEE Trans. Evol. Comput.* (2023), 1–1.

[71] B. Pes, Learning from high-dimensional and class-imbalanced datasets using random forests, *Information* **12**(8) (2021).

[72] A.G. Putrada, M. Abdurohman, D. Perdana and H.H. Nuha, Machine learning methods in smart lighting toward achieving user comfort: A survey, *IEEE Access* **10** (2022), 45137–45178.

[73] S. Qiao, N. Han, F. Huang, K. Yue, T. Wu, Y. Yi, R. Mao and C.-A. Yuan, LMNNB: Two-in-One imbalanced classification approach by combining metric learning and ensemble learning, *Appl. Intell.* **52**(7) (2022), 7870–7889.

[74] B.S. Raghuwanshi and S. Shukla, Classifying imbalanced data using BalanceCascade-based kernelized extreme learning machine, *Pattern Anal. Appl.* **23**(3) (2020), 1157–1182.

[75] C.A. Rolón-González, R. Castañón-Méndez, A. Alarcón-Paredes, I. López-Yáñez and C. Yáñez-Márquez, Improving the performance of an associative classifier in the context of class-imbalanced classification, *Electronics (Switzerland)* **10**(9) (2021).

[76] K. Roy, M. Ahmad, K. Waqar, K. Priyaah, J. Nebhen, S.S. Alshamrani, M.A. Raza and I. Ali, An Enhanced Machine Learning Framework for Type 2 Diabetes Classification Using Imbalanced Data with Missing Values, *Complexity* **2021** (2021).

[77] J.A. Sáez, J. Luengo, J. Stefanowski and F. Herrera, SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering, *Inf. Sci.* **291** (2015), 184–203.

[78] Y. Sanguanmak and A. Hanskunatai, Auto-tuning of parameters in hybrid sampling method for class imbalance problem, in: *2016 International Computer Science and Engineering Conference (ICSEC)*, 2016, pp. 1–5.

[79] Y. Sanguanmak and A. Hanskunatai, DBSM: The combination of DBSCAN and SMOTE for imbalanced data classification, in: *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, IEEE, 2016.

[80] A. Sbei, K. ElBedoui, W. Barhoumi and C. Maktouf, Adaptive feature selection in PET scans based on shared information and multi-label learning, *Visual Comput.* **38**(1) (2022), 257–277.

[81] H. Shamsudin, U.K. Yusof, Y. Haijie and I.S. Isa, AN OPTIMIZED SUPPORT VECTOR MACHINE WITH GENETIC ALGORITHM FOR IMBALANCED DATA CLASSIFICATION, *Jurnal Teknologi* **85**(4) (2023), 67–74.

[82] S.S. Shaw, S. Ahmed, S. Malakar, L. Garcia-Hernandez, A. Abraham and R. Sarkar, Hybridization of ring theory-based evolutionary algorithm and particle swarm optimization to solve class imbalance problem, *Complex & Intelligent Systems* **7**(4) (2021), 2069–2091.

[83] M.J. Siers and M.Z. Islam, Class Imbalance and Cost-Sensitive Decision Trees: A Unified Survey Based on a Core Similarity, *ACM Trans. Knowl. Discov. Data* **15**(1) (2021).

[84] A. Singh, A. Jain and S.E. Biable, Financial Fraud Detection Approach Based on Firefly Optimization Algorithm and Support Vector Machine, *Applied Computational Intelligence and Soft Computing* **2022** (2022).

[85] P. Soltanzadeh, M.R. Feizi-Derakhshi and M. Hashemzadeh, Addressing the class-imbalance and class-overlap problems by a metaheuristic-based under-sampling approach, *Pattern Recognit.* **143** (2023), 109721.

[86] V. Stanovov, E. Semenkin and O. Semenkina, Self-configuring hybrid evolutionary algorithm for fuzzy imbalanced classification with adaptive instance selection, *J. Artif. Organs* (2016).

[87] R. Storn and K. Price, Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces, *J. Global Optimiz.* **11**(4) (1997), 341–359.

[88] B. Sun, H. Chen, J. Wang and H. Xie, Evolutionary under-sampling based bagging ensemble method for imbalanced data classification, *Front. Comput. Sci.* **12**(2) (2018), 331–350.

[89] C. Sun, L. Ippel, A. Dekker, M. Dumontier and J. van Soest, A systematic review on privacy-preserving distributed data mining, *Data Sci.* **4**(2) (2021), 121–150.

[90] S. Susan and A. Kumar, The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art, *Eng. Rep.* (2020).

[91] E.F. Swana, W. Doorsamy and P. Bokoro, Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset, *Sensors* **22**(9) (2022).

[92] J.J. Tanimu, M. Hamada, M. Hassan, H. Kakudi and J.O. Abiodun, A Machine Learning Method for Classification of Cervical Cancer, *Electronics* **11**(3) (2022), 463.

[93] Y. Tao, B. Jiang, L. Xue, C. Xie and Y. Zhang, Evolutionary synthetic oversampling technique and cocktail ensemble model for warfarin dose prediction with imbalanced data, *Neural Comput. Appl.* **33**(17) (2021), 11203–11221.

[94] S.C.K. Tékouabou, C. Gherghina, H. Toulni, P.N. Mata and J.M. Martins, Towards Explainable Machine Learning for Bank Churn Prediction Using Data Balancing and Ensemble-Based Methods, *Sci. China Ser. A Math.* **10**(14) (2022).

[95] O. Ussatova, A. Zhumabekova, Y. Begimbayeva, E.T. Matson and N. Ussatov, Comprehensive DDoS Attack Classification Using Machine Learning Algorithms, *Computers, Materials and Continua* **73**(1) (2022), 577–594.

[96] M.E. Villa-Pérez, M.Á. Álvarez-Carmona, O. Loyola-González, M.A. Medina-Pérez, J.C. Velazco-Rossell and K.-K.R. Choo, Semi-supervised anomaly detection algorithms: A comparative summary and future research directions, *Knowledge-Based Systems* **218** (2021).

[97] A.V. Vitianingsih, Z. Othman, S.S. Kama, A. Suraji and A.L. Maukar, Application of the synthetic over-sampling method to increase the sensitivity of algorithm classification for class imbalance in small spatial datasets, *Int. J. Intell. Eng. Syst.* **15**(5) (2022), 676–690.
[98] S. Vluymans, I. Triguero, C. Cornelis and Y. Saeys, EPRENNID: An evolutionary prototype reduction based ensemble for nearest neighbor classification of imbalanced data, *Neurocomputing* **216** (2016), 596–610.
[99] X. Wang, Y. Li, J. Zhang, B. Zhang and H. Gong, An oversampling method based on differential evolution and natural neighbors, *Appl. Soft Comput.* **149** (2023), 110952.
[100] Z. Wang and Q. Liu, Imbalanced Data Classification Method Based on LSSASMOTE, *IEEE Access* **11** (2023), 32252–32260.
[101] Y. Xiao and M. Watson, Guidance on Conducting a Systematic Literature Review, *Journal of Planning Education and Research* **39**(1) (2019), 93–112.
[102] P. Yang, L. Xu, B.B. Zhou, Z. Zhang and A.Y. Zomaya, A particle swarm based hybrid system for imbalanced medical data sampling, *BMC Genomics* **10 Suppl 3**(Suppl 3) (2009), S34.
[103] L.P. Yulianti and K. Surendro, Implementation of quantum annealing: A systematic review, *IEEE Access* **10** (2022), 73156–73177.
[104] J. Zhang and I. Mani, kNN approach to unbalanced data distributions: a case study involving information extraction, Accessed: 2023-9-13.
[105] J. Zhang, Y. Li, B. Zhang, X. Wang and H. Gong, A new oversampling approach based differential evolution on the safe set for highly imbalanced datasets, *Expert Syst. Appl.* **234** (2023), 121039.
[106] Z. Zhang, S. Zhang, C. Chen and J. Yuan, A systematic survey of air quality prediction based on deep learning, *Alex. Eng. J.* **93** (2024), 128–141.
[107] Z. Zhang and J. Li, Synthetic Minority Oversampling Technique Based on Adaptive Local Mean Vectors and Improved Differential Evolution, *IEEE Access* **10** (2022), 74045–74058.
[108] Z.-L. Zhang, R.-R. Peng, Y.-P. Ruan, J. Wu and X.-G. Luo, ESMOTE: an overproduce-and-choose synthetic examples generation strategy based on evolutionary computation, *Neural Comput. Appl.* **35**(9) (2023), 6891–6977.
[109] Y. Zhu, Y. Yan, Y. Zhang and Y. Zhang, EHSO: Evolutionary Hybrid Sampling in overlapping scenarios for imbalanced learning, *Neurocomputing* **417** (2020), 333–346.
[110] R. Zuech, J. Hancock and T.M. Khoshgoftaar, Detecting web attacks using random undersampling and ensemble learners, *Journal of Big Data* **8**(1) (2021).