Artificial Intelligence Review (2024) 57:273
https://doi.org/10.1007/s10462-024-10884-2

Handling imbalanced medical datasets: review of a decade of research

Mabrouka Salmi[1,2] · Dalia Atif[3] · Diego Oliva[4] · Ajith Abraham[5] · Sebastian Ventura[2]

Accepted: 25 July 2024 / Published online: 2 September 2024

Abstract

Machine learning and medical diagnostic studies often struggle with the issue of class imbalance in medical datasets, complicating accurate disease prediction and undermining diagnostic tools. Despite ongoing research efforts, specific characteristics of medical data frequently remain overlooked. This article comprehensively reviews advances in addressing imbalanced medical datasets over the past decade, offering a novel classification of approaches into preprocessing, learning levels, and combined techniques. We present a detailed evaluation of the medical datasets and metrics used, synthesizing the outcomes of previous research to reflect on the effectiveness of the methodologies despite methodological constraints. Our review identifies key research trends and offers speculative insights and research trajectories to enhance diagnostic performance. Additionally, we establish a consensus on best practices to mitigate persistent methodological issues, assisting the development of generalizable, reliable, and consistent results in medical diagnostics.

Keywords Class imbalance · Medical datasets · Medical diagnosis · Machine learning

✉ Sebastian Ventura
sventura@uco.es

Mabrouka Salmi
z12salsm@uco.es

Dalia Atif
atif.dalia@cu-tipaza.dz

Diego Oliva
diego.oliva@cucei.udg.mx

Ajith Abraham
ajith.abraham@ieee.org

[1] Laboratory of Applied Statistics (LASAP), National Higher School of Statistics and Applied Economics, Koléa, Tipaza, Algeria

[2] Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Cordoba, Cordoba, Spain

[3] Economics Department, University Center of Tipaza, Tipaza, Algeria

[4] Depto. de Ingeniería Electro-Fotónica, Universidad de Guadalajara, CUCEI, Guadalajara, Jalisco, Mexico

[5] School of Artificial Intelligence, Bennett University, Greater Noida 201310, Uttar Pradesh, India

1 Introduction

The class imbalance issue remains one of the main challenges in data mining. The fact that one class is underrepresented in a dataset while the other(s) is prevailing results in uneven class distribution. When the data is unevenly distributed, the prevalent class is called the majority class, while the one containing the rare cases is called the minority class. The minority class is usually ignored by the machine learning algorithms that prioritize the majority class (Sun et al. 2009). Due to the imbalance in the dataset, conventional machine learning algorithms are biased towards the class primarily present in the data, while those rare cases are neglected. The main reason for such a problem is how machine learning algorithms are constructed; they assume balanced datasets (Krawczyk 2016). The balance in real-world datasets is often unreached, and the ill-prepared machine learning algorithms cannot assist in detecting rare cases of interest, which is an immense concern in research.

Medical diagnosis data are becoming of great use and interest with the progress in big data and medicine (Haixiang et al. 2017). Hence, it is subject to improving medical care treatment and creating aid-medical diagnosis systems. Machine learning is availing in designing medical diagnosis systems (Huda et al. 2016; Xiao et al. 2021; Woźniak et al. 2023); however, the imbalanced medical data hinders the machine learning algorithms' performance, thus the performance of medical diagnosis systems. Medical diagnosis data could be represented in two classes one of the non-diseased individuals (healthy) and the other of the diseased individuals (unhealthy). Accurately predicting unhealthy individuals (diseased patients) on time allows early access to medical treatment and saves patients' lives, which is unachieved without appropriate handling of class imbalance in medical datasets.

Intensive research has been conducted through literature to deal with the issue of class imbalance in general. Consequently, several methods of learning from imbalanced data have been proposed, and they are grouped mainly into two approaches: data-level and algorithmic level. The latter modifies the learning algorithms to consider the minority class, and the former handles the class imbalance by modifying the data distribution, whether through undersampling that eliminates instances from the majority class, oversampling the minority class that creates synthetic instances, or hybridizing both under and oversampling to reduce the imbalance. In addition, researchers propose several basic and advanced class imbalance handling methods that are generally applied to various domains.

Many literature reviews have been carried out on class imbalance, whether focusing on class imbalance handling methods only (Galar et al. 2011; Abd Elrahman and Abraham 2013; Spelmen and Porkodi 2018; Ali et al. 2019), both methods and applications (Haixiang et al. 2017; Kumar et al. 2021), or methods for a specific application's field (Patel et al. 2020). However, class imbalance in medical diagnosis is not well highlighted, yet specificities of the imbalanced medical data are unconsidered. Such specificities pose a unique challenge for working with medical data and require specialized techniques and methodologies to ensure the validity and generalizability of the findings. Improving existing medical diagnosis systems and improving human well-being rely on medical diagnosis research. Hence, researchers and practitioners in healthcare, in general, and in medical diagnosis need to be aware of these factors and be abreast of the recent advancements in the field to identify their starting research points. In this work, we analyze the literature on handling imbalanced medical datasets and formulate the following intended research questions to cover the knowledge gaps.

- RQ1 How can we develop a comprehensive framework for categorizing and evaluating imbalanced learning techniques tailored specifically to the complexities of medical datasets?
- RQ2 What emerging trends and future trajectories are envisaged for tackling imbalanced medical data?
- RQ3 What methodological techniques and procedural recommendations for mitigating class imbalance in research studies with a focus on enhancing the validity and reliability of results?

We aim to emphasize the research on the intersection of class imbalance in structured data and medical diagnosis through a well-designed research methodology. This paper comprehensively reviews the last decade's research and clusters the reviewed literature in medical imbalanced datasets in three main approaches by building up on the existing classification of class imbalance methods (Krawczyk 2016): preprocessing level entailing data level methods and feature level methods, learning level encloses algorithmic methods, and combined techniques hybridize the two mentioned approaches. Related research is meticulously classified into subgroups within each approach to specifically present the state-of-the-art and facilitate detailed tracking of advancements and areas for continued development. This review systematically extracts and presents detailed statistics on the medical datasets and evaluation metrics employed in existing literature, delineating the most and least commonly used resources to offer insights into prevailing research methodologies. It synthesizes prior research outcomes concerning class imbalance in medical datasets and discusses observations from the contextual analysis. This innovative exploration offers speculative insights into methodological concerns and practical aspects, critically evaluating the high performance of specific methodologies across diverse medical datasets. Subsequently, we acknowledge the inevitable limitations of our study due to non-reproducible experimental outcomes and other significant constraints encountered in the analysis of imbalanced medical data. In addition to presenting original contributions, this review identifies research trends in imbalanced medical datasets and highlights promising directions for future research that could enhance medical diagnosis performance. It also establishes best practices in this field, aiming to mitigate prevalent issues and proposing a consensus among researchers to guide future studies.

The structure of the review paper is as follows: Sect. 2 introduces the problem of class imbalance in medical datasets. Section 3 details the search methodology and describes the findings regarding used medical datasets and evaluation metrics. Section 4 presents the data-level approach proposed for imbalanced medical datasets, Sect. 5 exposes the learning-level proposed solutions, and Sect. 6 contains the proposed combined techniques in the literature. Section 7 synthesizes the outcomes of research works on several imbalanced medical datasets. Section 8 discusses reflections on the synthesis, highlighting speculative insights, whereas the value and limitation of the observatory synthesis are pointed out in Sect. 9. Section 10 summarizes the research trends and future directions in imbalanced medical datasets research. In Sect. 11, we highlighted the best practices amongst researchers in imbalanced medical data. Section 12 concludes the paper.

2 The problem of class imbalance in medical data

With the advancement of technologies, medical data is increasingly stored in the form of electronic medical records, where the historical medical data of an individual is saved and shared with authorized users (Fujiwara et al. 2020). Demographic data, clinical tests, X-ray images, MRIs, fMRI, EEG, and other types represent medical information. The access to voluminous medical data, along with the progress in the application of machine learning, has been helpful for medical care specialists and clinicians. Machine learning effectiveness in multiple domains encourages constructing aid-medical diagnosis systems to automate medical diagnosis and help with the scarcity of medical experts in specific domains and places and the vast demand for diagnosis for specific diseases. Those diagnosis systems are trained on historical medical data about a particular disease to perform well on unknown new medical data and predict the disease. However, such systems are constructed through well-designed processes depending on the disease and its data availability with the help of experts' knowledge. Nonetheless, the class imbalance in medical data hardens the mission of machine learning algorithms and diagnostic systems.

While naturally unhealthy people are less than healthy, the class imbalance exists if the classes are unequally distributed in the dataset for training machine learning algorithms. There are numerous sources of imbalance in medical data. However, they can be grouped into four patterns:

- *Bias in data collection*: resulting from the fact that certain groups, such as non-diabetics, are underrepresented in research because they are underdiagnosed.
- *The prevalence of rare classes*: in this case, the imbalance is inherent to the disease because certain conditions occur in 1 per 100,000 in the population, making the positive class rare.
- *Longitudinal studies*: medical studies investigated over time can result in an imbalance in the dataset due to the discharge of certain patients (lost to follow-up) or the change of class over time (such as the progression of one stage to another in the case of cancer).
- *Data privacy and ethics*: the susceptibility of certain diseases, such as HIV, can limit access to positive classes, resulting in imbalanced datasets.

An imbalanced dataset is defined by a disproportionate distribution between classes, where the Imbalance Ratio (IR), calculated as $IR = N_{maj}/N_{min}$, indicates the extent of this disproportion. In this formula, N_{maj} and N_{min} represent the number of instances in the majority and minority classes, respectively. In binary datasets, the degree of imbalance is usually defined as $IR : 1$, where the more significant the difference than 1, the more severe the imbalance is.

Many existing classifiers exhibit an inductive bias that favors the majority class when trained on imbalanced datasets, often at the expense of the minority class. This results in suboptimal performance in less-represented classes. For instance, in diagnoses such as cancer risk or Alzheimer's disease, patients are typically outnumbered by healthy individuals. Unfortunately, conventional classifiers tend to prioritize high overall accuracy, potentially leading to the misclassification of at-risk patients as healthy. Such errors in classification can have grave consequences, including the inappropriate discharge of patients in need of critical care. Additionally, this predisposition can lead to unfair treatment and ethical dilemmas, as it systematically disadvantages those requiring the most medical attention, raising significant concerns about equity in healthcare diagnostics.

Class imbalance handling methods are created for general purposes and not for medical diagnosis data. Applying such methods without considering the context of the disease in matter or the data at hand may lead to uninterpretable yet inaccurate results (Han et al. 2019). For example, synthetic minority data are generated to balance the medical data so machine learning can learn equally on both existing classes (diseased patients and non-diseased patients). However, synthetic data needs to conform to the characteristics of original medical data. Besides, the application of machine learning algorithms for medical diagnosis needs to be adequately evaluated in case of imbalanced data. The cost of misclassifying a diseased patient is more critical than misclassifying a non-diseased patient. The first can lead to dangerous consequences that may affect the patient's life, whereas the second may lead to a further clinical investigation (Fotouhi et al. 2019). Therefore, the evaluation of medical diagnosis machine learning models relies mainly on measuring their predictive power for minority cases (diseased patients) (Han et al. 2019). However, a well-performing medical diagnosis system is expected to provide the best compromise in predicting diseased and non-diseased patients and avoid all kinds of costs of misclassification.

On the other hand, synthetic data must adhere to the original medical data's characteristics. Otherwise, the automatic application of generic methods such as SMOTE (Chawla et al. 2002) may introduce biases and patterns not present in the original data, as well as irrelevant biologically impossible information, which may affect overall model performance for a variety of reasons, including inaccurate representation of rare case characteristics leading to unreliable model predictions, creation of synthetic data only in the rare cases neighborhood causing overfitting and generalization problems, and worst feature representation by increasing, decreasing, or reversing a variable's impact on the target. Researchers have thus worked over the last decade to find solutions that avoid the drawbacks mentioned earlier, such as creating synthetic instances more representative of the underlying distribution, reducing the risk of inducing noise, and ensuring better generalization. Misdiagnosis occurs due to the difficulty in learning rare cases, and the need for researchers to stay up-to-date with the latest advances motivates them to incorporate improvements in the field into their research to maximize the utility of available data. As a result, our motivation in this paper is to classify pertinent techniques into several strata and to provide a critical review of the relevant literature as well as a synthesis of the outcomes of research on reference class imbalance datasets based on several metrics, enriching this classification to enhance the advantages and disadvantages of each stratum, thus opening up new research directions in the field.

3 Research methodology and basic statistics

This section details the search methodology used for data collection and the statistics describing the extracted data from the reviewed literature. The proposed review process follows most of the common guidelines proposed by Kitchenham (2004) for performing systematic literature reviews in software engineering research.

3.1 Data collection

Our search methodology defined the bibliographic databases, search keywords, inclusion and exclusion criteria, and time range for our literature review. Regarding the bibliographic databases, we selected Google Scholar and Scopus to collect papers. Besides, the search keywords are shown in Fig. 1. The inclusion and exclusion criteria used for paper selection and the search methodology process are illustrated in the diagram (see Fig. 2).

We used advanced search in the Google Scholar and Scopus databases to find papers, setting the search for keywords in title only with a time range between 2013 and 14/01/2023. In our search, we used keywords to search for papers with class imbalance as a topic like "imbalanced" and keywords to capture papers that treated medical data in general like "medical" as depicted in Fig. 1. In addition, we eliminated some search terms due to their widespread occurrence with search keywords like "diagnosis", based on some trials, and to ensure the relevance of the results. The preliminary results were 409 in Google Scholar and 222 in Scopus, which we added to our reference manager. The first cleaning of our collected dataset by removing the duplicates and some unrelated results ended up with 249 papers. Afterward, we scanned the collected papers through the title and abstract if needed to sort out only relevant papers according to our review scope and the selection criteria. This second scanning yielded 165 papers pertinent to our review topic. A final scanning of the remaining results through full text was necessary, and we ended up with 150 papers, among them twelve without access. For that, the reviewed papers in this article are 137. The diagram in Fig. 2 illustrates the proposed methodology for data collection.

3.2 Analyses of used datasets and classification-based metrics

3.2.1 Medical datasets

We extracted all the datasets used in the reviewed research articles and grouped them based on their availability: public or private. We found that 95 (69%) papers used publicly available medical datasets, and 44 papers used private ones. Some research articles use both public and private datasets, and three research papers could have mentioned their employed datasets more clearly. The public datasets used in research have been investigated by extracting their usage frequency. Therefore, those datasets are partitioned into two groups: reference class imbalance medical datasets, which are frequently used (see Fig. 4), are displayed in Table 1, and non-reference class imbalance medical datasets are less used than

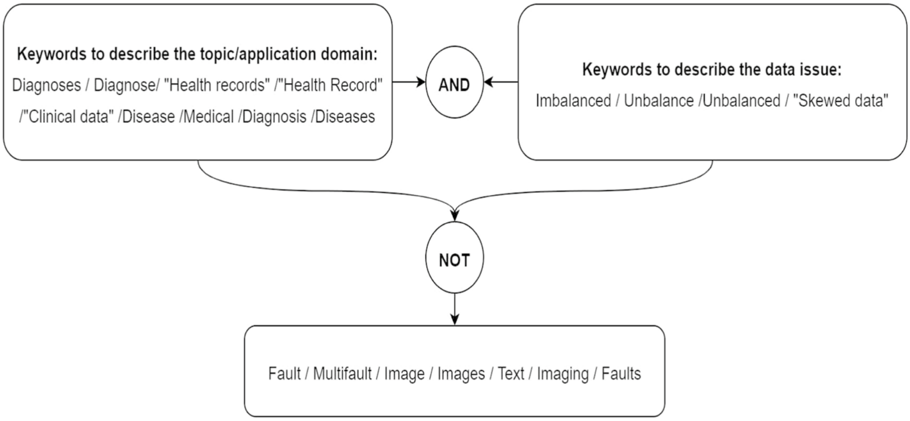
Keywords to describe the topic/application domain:
Diagnoses / Diagnose/ "Health records" /"Health Record" /"Clinical data" /Disease /Medical /Diagnosis /Diseases
AND
Keywords to describe the data issue:
Imbalanced / Unbalance /Unbalanced / "Skewed data"
NOT
Fault / Multifault / Image / Images / Text / Imaging / Faults

Fig. 1 The used search keywords

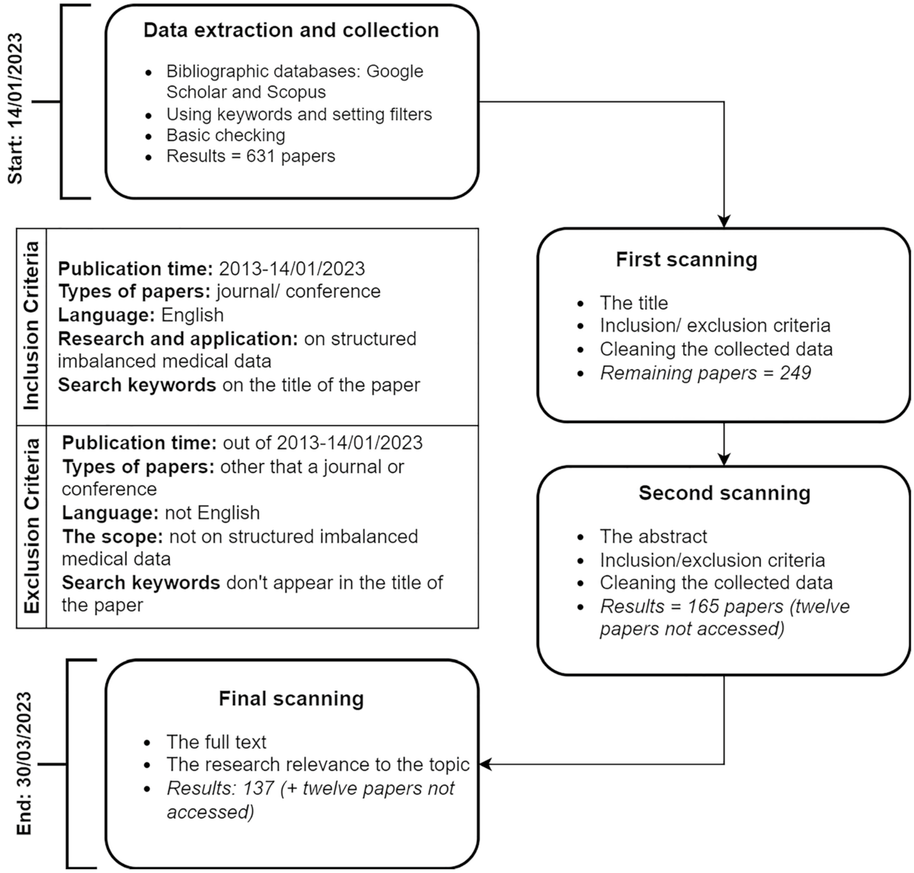
Start: 14/01/2023
Data extraction and collection
• Bibliographic databases: Google Scholar and Scopus
• Using keywords and setting filters
• Basic checking
• Results = 631 papers
Inclusion Criteria
Publication time: 2013-14/01/2023
Types of papers: journal/ conference
Language: English
Research and application: on structured imbalanced medical data
Search keywords on the title of the paper
Exclusion Criteria
Publication time: out of 2013-14/01/2023
Types of papers: other that a journal or conference
Language: not English
The scope: not on structured imbalanced medical data
Search keywords don't appear in the title of the paper
First scanning
• The title
• Inclusion/ exclusion criteria
• Cleaning the collected data
• Remaining papers = 249
Second scanning
• The abstract
• Inclusion/exclusion criteria
• Cleaning the collected data
• Results = 165 papers (twelve papers not accessed)
End: 30/03/2023
Final scanning
• The full text
• The research relevance to the topic
• Results: 137 (+ twelve papers not accessed)

Fig. 2 The search methodology for the literature review

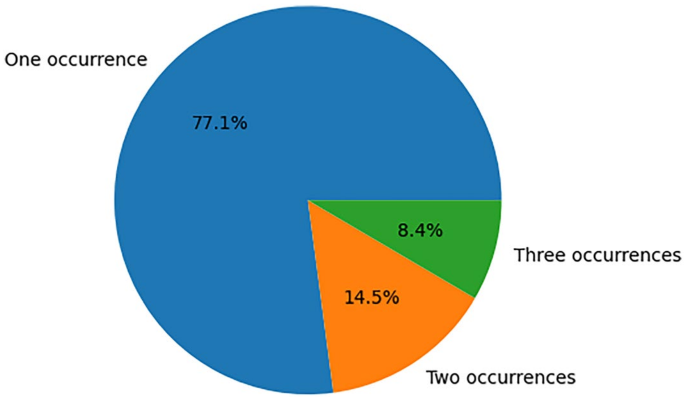

Fig. 3 Non-reference class imbalance medical datasets

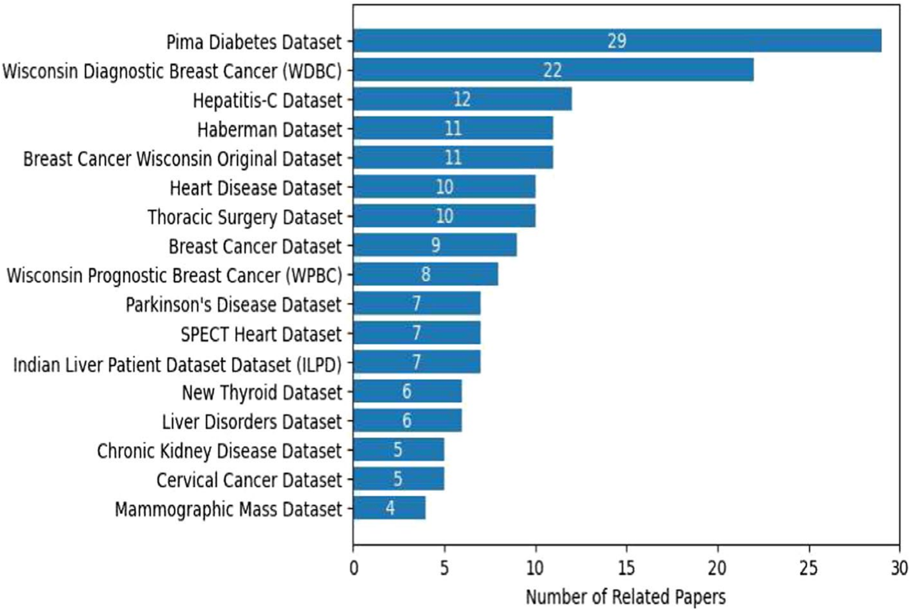

Fig. 4 Reference class imbalance medical datasets

those above. Figure 3 illustrates the non-reference datasets based on their commonness in research.

Table 1 displays the main characteristics of the medical datasets as used in the reviewed research, including the dataset size, the number of features, and the number of instances in each of the minority and the majority classes. All the medical datasets are initially of binary class except for the "New Thyroid Disease Dataset," which consists of three classes. For deeper insights into the procedural and contextual specifics of the dataset, it is advised to refer to the detailed discussions found in the referenced data sources and the foundational studies. The imbalance varies from one dataset to another, indicating the difference in its degree; while a dataset is highly imbalanced in research work, it is moderately imbalanced in another research framework; a class imbalance of one dataset is slight compared to another but could be more challenging. Although this points to the lack of an accepted universal quantification of the severity degree of the imbalance - as discussed later in Sect. 8, the imbalance of the datasets in Table 1 is highlighted and well-considered as an imbalance across the literature. While reviewing the reference medical datasets, we identified an underrepresentation of certain medical domains, such as psychiatry and psychology. This absence may be linked to the data scarcity as stated by Kumar et al. (2023), or the nature of these fields, which are often explored through unstructured, text-based data (Awon et al. 2022), thus falling outside the primary scope of our structured data analysis.

3.2.2 Evaluation metrics and statistical tests

Reviewed research papers selected from different evaluation metrics to assess the performance of their proposed approach. Several metrics and statistical tests have been used in medical diagnosis using imbalanced datasets. We extracted all the used metrics and statistical tests in the reviewed literature and presented the findings in Fig. 5 and Table 2.

Table 1 Description of reference class imbalance medical datasets

Dataset	Number of instances	Number of features	Minority class count	Majority class count	Number of classes	Imbalance ratio
Breast cancer Wisconsin original dataset	699	10	241	458	2	1.90:1
Heart disease dataset	303	76	138	165	2	1.20:1
Cervical cancer dataset	858	36	55	803	2	14.6:1
Hepatitis-C dataset	155	19	32	123	2	3.84:1
Indian Liver patient dataset (ILPD)	583	10	167	416	2	2.49:1
Breast cancer dataset	286	9	85	202	2	2.38:1
SPECT heart dataset	267	22	55	212	2	3.85:1
Haberman dataset	306	4	81	225	2	2.78:1
Wisconsin prognostic breast cancer (WPBC)	198	34	47	151	2	3.21:1
Wisconsin diagnostic breast cancer (WDBC)	569	32	212	357	2	1.68:1
Pima diabetes dataset	768	9	268	500	2	1.87:1
Parkinson’s disease dataset	197	23	48	147	2	3.06:1
New thyroid disease dataset	215	5	35	180	2	5.14:1
Chronic kidney disease dataset	400	25	150	250	2	1.67:1
Thoracic surgery dataset	470	17	70	400	2	5.71:1
Liver disorders dataset	345	7	145	200	2	1.38:1
Mammographic mass dataset	961	6	445	516	2	1.16:1

Imbalance Ratios are calculated as the ratio of the majority class size to the minority class size for each dataset (N_{maj}/N_{min}). A ratio significantly greater than 1 indicates a higher degree of imbalance, with the data becoming increasingly imbalanced as the ratio deviates further from 1

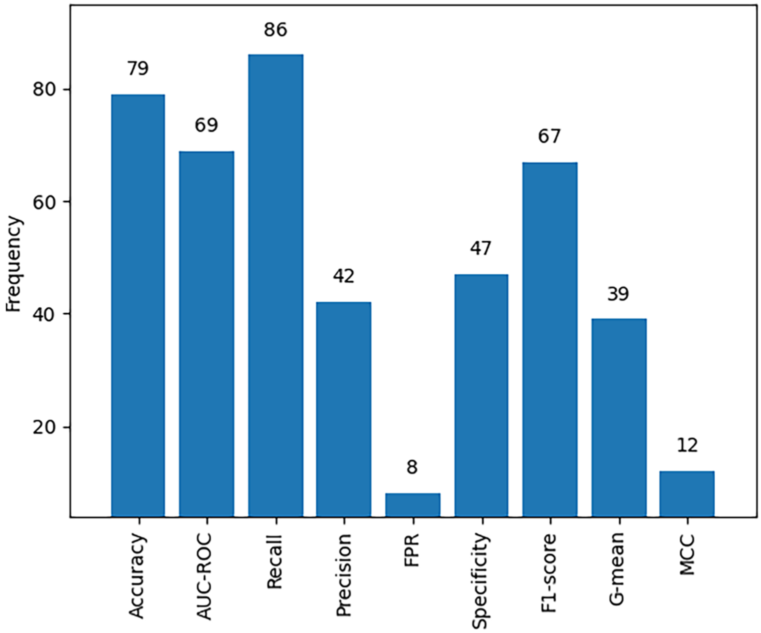

Fig. 5 Frequently used metrics

Table 2 Infrequently used metrics and statistical tests

Usage frequency	Metric
1	Modified accuracy; recall to majority class; precision of majority class; recall to minority class; precision of minority class; brier skill score; average of sensitivity and specificity; average precision; Kolmogorov–Smirnov distance; holm test; pairwise *t*-test; Nemenyi test; Fowlkes–Mallows (FMI); mean absolute error; maximum geometry mean (MGM); relative absolute error; Kruskall Wallis variance analysis test; *k*-value; covariance; average aggregated metric; Tukey's honest significant difference (HSD); deviance information criterion (DIC); the expected information criteria (Akaike (EAIC); the Schwarz and Bayesian information criteria (EBIC); Watanabe–Akaike information criterion (WAIC); the Hannan–Quinn information criterion (HQIC); the Bayesian predictive information criterion (or simply IC); error rate; recall macro; precision macro; loss function; f1-score for the minority class; average f1; averaged accuracy; G-mean for the minority class
2	Hamming loss; Anova
3	Macro-F1; RMSE
4	NPV; Wilcoxon paired signed rank; TN
5	FNR; Kappa; Friedman test; TP; balanced accuracy; AUC-PR
6	PPV

The used metrics and statistical tests are split into two groups: the first group contains frequently used metrics used at least eight times in the literature. In contrast, the second group contains infrequently used both metrics and statistical tests used a maximum of seven times.

As seen in Fig. 5, nine classification-based metrics are primarily used: AUC-ROC, Recall (also known as sensitivity), Precision, Specificity, F1-score, G-Mean, FPR (False positive ratio), Matthews Correlation Coefficient (MCC), and Accuracy. We notice that recall is the most used metric. 62.8% papers selected the recall to assess their proposed approaches. In the case of imbalanced data, the focus is on classifying minority classes, especially when dealing with imbalanced data in medical diagnosis. For that, using sensitivity is essential in class imbalance research. Furthermore, accuracy, AUC-ROC, and F1-score are used in medical diagnosis systems evaluation. Accuracy is used in 57% of the reviewed literature, while it reflects the overall performance of models and hides the misclassification of minority examples. Research emphasizes the use of recall (sensitivity) to measure the model's performance in identifying minority samples that are unhealthy or diseased patients in our case. However, we found that accuracy is still widely used in second place after sensitivity and is used solely to evaluate the imbalanced classification model in several studies (Sajana and Narasingarao 2018b; Mohd et al. 2019; Babar 2021; Lan et al. 2022). The area under the curve, also known as the AUC-ROC or AUC, is significantly used in 50.4% of the explored literature. The information added by the area under the curve indicates the ability of the proposed approach to discriminate between the minority class and the majority class. The higher the value of the AUC, the more powerful the model is in discrimination between different classes. Therefore, we notice the importance attributed to the AUC in medical diagnosis with imbalanced data research, where some researchers rely individually on it in experimental analysis (Çinaroğlu 2017; Hassan and Amiri 2019). Another commonly used metric is the *F*-value, used in 49% of reviewed literature. The *F*-value informs about balanced classification; the higher its value, the better the trade-off between precision and recall. Referring to the *F*-value, the misclassified minority examples and the misclassified majority examples are considered. As a result, the model performance in classifying both classes in binary classification is evaluated by the *F*-value. The high frequency of using the *F*-value indicates the attention to both minority and majority classes, hence the general performance of proposed approaches in handling imbalanced medical datasets.

Another cluster of quietly used metrics contains specificity, precision, and geometric mean (GM). 34% papers utilized specificity, while 30.7 and 28% of included literature used precision and geometric mean, respectively. The precision metric focuses on minimizing the false minority predicted examples, for that reference to it is vital to the excellent performance of medical diagnosis models; however, it is not as important as recall. Focusing only on recall minimizes the type one error in classification. Hence, we avoid predicting a diseased patient as non-diseased, and an accurate diagnosis saves human lives by allowing patients to access treatment as early as possible. However, by ignoring and minimizing the false predicted minority examples non-diseased patients are diagnosed as diseased patients, a type II error in classification, which may charge extra costs for all parts of society (like medical care and patients). The degree of attention to each classification error may be the reason for the difference in recall and precision in medical diagnosis research. Specificity and G-mean are used frequently but less frequently than other metrics like recall and accuracy. In some research works (Naghavi et al. 2019; Liu et al. 2020; Ibrahim 2022), we see that selecting one metric like specificity or G-mean or both with the recall to evaluate the proposed approaches. The G-mean of sensitivity and specificity shows the compromise between both metrics. When used with sensitivity, it can inform about the specificity score. Besides, the specificity quantifies the model's ability to identify the majority class; knowing the specificity aside from the sensitivity illustrates the balance between them. Consequently, the relatively lower use of specificity and G-mean compared to recall is mainly explained as mentioned. However, we see considerable attention to recall in other research works without referring to the specificity and G-mean (Sun et al. 2021; Mienye and Sun 2021; Shi et al. 2022), which ignore the balance between both correctly identifying diseased patients as well as correctly identifying non-diseased patients.

Moreover, the Matthews correlation coefficient, also known as MCC, that informs about the classification performance could be even better than the *F*-value, and accuracy (Xu et al. 2020) is less frequently used. The False positive rate (FPR) is considered even though less than other metrics 9% papers used it. It refers to the misdiagnosed cases as diseased individuals. Thus, we notice a growing convergence of researchers to other metrics to quantify the performance of their proposed disease diagnosis models Accuracy, *F*-value instead of MCC, and FPR instead of sensitivity. Some research works used them simultaneously with standard metrics to better analyze the model's effectiveness (Shilaskar et al. 2017; Sadrawi et al. 2018; Cheng and Wang 2020).

Table 2 groups uncommonly used metrics and statistical tests with their frequency of usage in reviewed research. We notice that statistical tests like the Friedman test, Wilcoxon paired signed rank, and Holm test are being used, even occasionally, which means researchers are referring to other tools, unlike evaluation metrics, to compare between proposed approaches and existing approaches. We find that the area under the precision-recall curve, also known as AUC-PR, is used only six times, although it is known as an appropriate metric for imbalanced classification (Huo et al. 2022; Albuquerque et al. 2022). A high AUC-PR means high precision and recall; therefore, it summarizes the model's predictive power in minority and majority classes. Other evaluation metrics are used in a few studies, and the necessity of some adapted metrics to the proposed models may explain the variety of used metrics and change their interpretation.

4 Pre-processing level

4.1 Feature level

It entails all methods focusing on feature space to treat class imbalance in the data. One of the existing feature-level methods is feature selection, a widely used preprocessing procedure in different machine-learning tasks that employs various techniques to retain discriminating features. Another feature-level method is feature extraction, which creates new features from the initial feature space to keep most information in a smaller new set of features. Both methods, generally, are used to deal with high dimensional data, where the selected or extracted features are supposed to be informative features and facilitate the learning process and model generalization. Alternatively, feature weighting is found in the literature to improve recognition of the class of interest that is usually rare in medical applications such as medical diagnosis and risk prediction. Recently, mentioned methods are proposed to handle imbalanced learning, whether as self-standing approaches (Zhang and Chen 2019a; Li et al. 2022; Shakhgeldyan et al. 2020), discussed in this section, or combined with other class imbalance techniques (Wang et al. 2020; Tang et al. 2021; Lijun et al. 2018).

Feature-level methods are used to tackle the class imbalance and reduce the dimensionality (Zhang and Chen 2019a; El-Baz 2015; Sridevi and Murugan 2014; Li et al. 2022). Researchers in Zhang and Chen (2019a) selected the optimal features of the breast tumor using an improved Laplacian score (LS), which better compromised computational efforts and classification performance by surpassing rough set-EKNN (El-Baz 2015) and feature selection-multiple layer perceptron (FSMLP) (Sridevi and Murugan 2014). Similarly, in Li et al. (2022), insightful selection of interpretable features using functional principal component analysis on longitudinal data achieved more accurate data categorization and reduced computing complexity. Filters and wrappers have been used in disease and mortality prediction, respectively (Venkatanagendra and Ussenaiah 2019; Shakhgeldyan et al. 2020). In comparison, feature selection using filters improved the classification performance of Feed-Forward NN, SVM, XG Boost, Random Forest, and LDA in Venkatanagendra and Ussenaiah (2019). A four-stage feature selection based on filter and wrapper methods exceeded random forest and logistic regression in Shakhgeldyan et al. (2020). Promisingly feature weighting yielded high discrimination between majority and minority data (Polat 2018; Baniasadi et al. 2020). Polat used similarity and clustering considering the class label to weight each attribute's data points, making them more linearly separable and illustrating superior results than random subsampling (Polat 2018). Baniasadi et al. applied linear interpolation for missing values imputation and sample weighting (Baniasadi et al. 2020). Feature-level methods are remarkably proposed once imbalanced data is highly dimensional. Unexpectedly, feature weighting provides promising results. It is necessary to investigate its efficiency in dealing with the class imbalance issue regardless of the high dimensionality. Table 3 briefly describes the feature-level methods.

4.2 Data level

This approach deals with class imbalance at the data level by modifying the data distribution to balance the dataset through oversampling, undersampling, or a combination. Oversampling augments the number of minority samples (rare cases) in the dataset using different techniques, and undersampling decreases the number of majority samples. Hybrid methods combine oversampling and undersampling to obtain evenly distributed data. Researchers commonly use data-level methods to address class imbalances due to their simple implementation in the preprocessing phase, which is independent of the learning process. In general, the versatility of resampling in imbalanced learning has been noticed earlier (Abd Elrahman and Abraham 2013; Haixiang et al. 2017); however, this section reviews their application in medical data.

4.2.1 Oversampling

Oversampling prevails in imbalanced medical data classification and is significantly referred to in assessing proposed class imbalance methods. Hereafter, oversampling is individually used to combat the imbalance issue.

Random oversampling with random forests showed optimal performance in identifying the severity of the Hepatitis C virus (Orooji and Kermani 2021). However, randomly duplicating original minority samples may lead to overfitting, which implies using advanced techniques. Popularly used technique SMOTE (Synthetic Minority Oversampling Technique) created by Chawla et al. (2002) outperformed with KNN classifier (Hassan and Amiri 2019), however, demonstrated similar results with logistic

Table 3 Feature-level methods for imbalanced medical data

In-text citation	Key idea	Year
Polat (2018)	Feature weighting based on similarity and clustering considering the class label	2018
Zhang and Chen (2019a)	Improved Laplacian score (LS) to select the optimal features of the breast tumor	2019
Venkatanagendra and Ussenaiah (2019)	Filters for feature selection	2019
Baniasadi et al. (2020)	Linear interpolation for missing value imputation and sample weighting	2020
Shakhgeldyan et al. (2020)	A four stages-based feature selection method combined with some machine learning classifiers	2020
Li et al. (2022)	A non-parametric classification method based on functional principal component analysis for feature extraction	2022

regression to threshold adjustment based on Youden index (YI) (Albuquerque et al. 2022). Recently, the data distribution of disease samples is emphasized in SMOTE oversampling (Xu et al. 2021; Sun et al. 2021). Xu et al. used SMOTE based on a filtered k-means clustering (KNSMOTE) to overcome noise generation, overlapping and borderline issues, which outpaced traditional and cluster-based oversampling (Xu et al. 2021). Sun et al. integrated a multi-dimensional Gaussian probability hypothesis test to add SMOTE synthesized samples (MDGPH-SMOTE) to the original minority samples, illustrating better classification accuracy and recall (Sun et al. 2021).

SMOTE was adapted to various data contexts and combined with machine learning algorithms (Mustafa et al. 2017; Wang et al. 2013; Mohd et al. 2019). Farther Distance-based SMOTE was used along with PCA to handle high dimensional imbalanced biomedical data, revealing superiority over correlation and information gain (Mustafa et al. 2017). Differently, Wang et al. structured a Minimum Spanning Tree based on the KNN graph for minority data, then SMOTE synthesized samples along the paths between two randomly selected samples (Wang et al. 2013). In multi-class medical data, SMOTE with MLP model attained the highest accuracy (Mohd et al. 2019). Sajana and Narasingarao (2018a), authors intentionally balanced the initial data with SMOTE then split it for learning and testing a Naive Bayes classifier. Researchers investigated the real class of artificial minority instances created by SMOTE (Sug 2016; Naseriparsa et al. 2020). Sug checked the class of synthetic data using MLP and accordingly trained tree classifiers; however, results revealed insignificant differences (Sug 2016). Generating synthetic samples within the region with a high density of minority samples reduced the class mixture (Naseriparsa et al. 2020) and exceeded SMOTE variants.

Alternatively, Oversampling-based diverse methods yielded positive results. Oversampling based on causal relationships between features exceeded CCR (combined cleaning and resampling algorithm), k-means SMOTE, GAN (Generative Adversarial Networks), and CUSBoost (cluster-based undersampling with boosting (Luo et al. 2021). Oversampling using improved ant colony to diagnose outpatients of TCM (Traditional Chinese medicine) exceeded traditional ML like C4.5. and SMOTE (Bi and Ma 2021).

The decomposition of minority data was extensively studied as a prior step to sampling (López et al. 2013; Napierala and Stefanowski 2016, 2012), yet no universal method was concluded. Han et al. (2019), the authors applied different sampling strategies based on minority data selection using a self-adaptive algorithm and enhanced the recognition of minority class. Very recent research investigated synthetic samples fitting in the minority data (Rodriguez-Almeida et al. 2022), unexpectedly, experiments revealed higher similarity between synthetic and real data did not necessarily improve the classification performance. Data generation-based deep learning approaches in structured data are emerging (Xiao et al. 2021; Lan et al. 2022). While GAN and SMOTE highly increased the classification accuracy in Lan et al. (2022), combining SMOTE variants with conditional tabular generative adversarial networks (CTGAN) yielded unstable results (Rodriguez-Almeida et al. 2022). In contrast, a Wasserstein generative adversarial network (WGAN) in gene expression data excelled popular sampling methods (Xiao et al. 2021).

Oversampling is relatively used on its own to treat class imbalance in disease prediction. Besides using existing oversampling techniques and combining or improving them, we see two recent lines of research. One that considers the data distribution and its specificities in medical diagnosis while sampling minority examples. The other line adopts generative adversarial networks in structured medical data, a newborn research topic, resulting in a hybridization of both lines as observed. However, both research topics are unexplored and

Table 4 Oversampling techniques

In-text citation	Key idea	Year
Wang et al. (2013)	SMOTE based on the Minimum Spanning Tree (MST) of the minority class samples	2013
Sug (2016)	MLP to check the label of SMOTE synthetic samples	2016
Mustafa et al. (2017)	The Farther Distance based SMOTE	2017
Sajana and Narasingarao (2018a)	SMOTE oversampling and Naïve Bayes for classification	2018
Mohd et al. (2019)	SMOTE oversampling for imbalanced multi-class medical dataset	2019
Han et al. (2019)	Oversampling based on the minority samples distribution	2019
Hassan and Amiri (2019)	Oversampling with SMOTE	2019
Naseriparsa et al. (2020)	Region-based SMOTE to overcome the class mixture drawback	2020
Luo et al. (2021)	Oversampling based on causal relationships between variables	2021
Bi and Ma (2021)	Improved ant colony for oversampling	2021
Orooji and Kermani (2021)	Random oversampling	2021
Xiao et al. (2021)	Wasserstein Generative Adversarial Networks (WGAN) model	2021
Xu et al. (2021)	A SMOTE using oversampling ratios based on filtered k-means clustering	2021
Sun et al. (2021)	Multi-dimensional Gaussian probability density hypothesis testing for SMOTE synthetic data (MDGPH-SMOTE)	2021
Rodriguez-Almeida et al. (2022)	Oversampling and data generation	2023
Albuquerque et al. (2022)	SMOTE and adjusting the threshold by maximizing the Youden index (YI)	2022
Lan et al. (2022)	SMOTE combined with Generative Adversarial Networks (GAN)	2023

open for investigation. Table 4 briefly describes the proposed oversampling techniques with their key ideas.

4.2.2 Undersampling

Undersampling decreases the number of prevalent class examples by removing noisy data or duplicates that are uninformative through basic techniques like random undersampling or advanced techniques like clustering-based ones. Although undersampling is less used than oversampling, it is inventively proposed in medical diagnosis research.

Random undersampling with Random Forest output superior performance in Covid-19 mortality prediction (Iori et al. 2022), Hereditary Angioedema disease diagnosis (Dai and Hua 2016), and melanoma prediction (Richter and Khoshgoftaar 2018). K-means clustering was integrated into undersampling and boosted the prediction of diseased patients (Augustine and Jereesh 2022; Neocleous et al. 2016; Babar and Ade 2016). Augustine & Jereesh balanced the data using random undersampling at the generated clusters level (Augustine and Jereesh 2022). While Neocleous et al. (2016) used k-nearest neighbours after clustering. Similarly, the authors in Babar and Ade (2016) designed a Multiple Linear Perceptron (MLPUS) using k-means clustering that outperformed SMOTE, where iteratively samples close to the cluster centroid were used to train MLP and only samples with the highest SM (Stochastic measure) values are added to the training data which keeps hard to learn samples. Simply, clustering the majority class into subsets equal to the minority class and combining each with the minority class for training modestly ameliorated the results in Li et al. (2018) and sometimes outperformed SMOTE in Rahman and Davis (2013). However, Ensembling base classifiers built on balanced subsets exceeded BalanceCascase and EasyEnsemble undersampling techniques (Parvin et al. 2013). Salman & Vomlel further weighted instances using mutual information at each cluster, and their trained Tree-Augmented Naive Bayes (TAN) surpassed TAN with SMOTE (Salman and Vomlel 2017). Recently, Ibrahim used Salp swarm optimization to efficiently determine the clusters' centres, which sometimes exceeded cluster-based sampling techniques (Ibrahim 2022).

Adding high-quality majority samples to the minority class is variedly suggested (Zhang et al. 2020; Wang et al. 2020). After randomly selecting a subsample of the majority samples, only those with high entropy were selected based on the Gaussian Naive Bayes estimator which hastened the undersampling process (Zhang et al. 2020). The results in Wang et al. (2020) significantly outpaced SMOTE and random undersampling using Imbalanced Self-Paced Learning (ISPL) with logistic regression. The authors in Al-Shamaa et al. (2020) separated majority class instances and minority class instances based on the Hellinger distance, and majority instances most similar to their neighbouring minority instances were added to the original minority class. Investigations showed higher performance of the method than Tomeklinks, random undersampling, and edited nearest neighbours.

The data distribution is distinctly integrated into undersampling (Vuttipittayamongkol and Elyan 2020b; Kamaladevi and Venkatraman 2021). Vuttipittayamongkol and Elyan (2020b) identified overlapped instances using recursive search neighbouring then discarded the majority class instances. While in Kamaladevi and Venkatraman (2021), the authors imputed noise samples using the mean and relabeled borderline samples based on Tversky similarity Indexive regression. Investigations illustrated promising results yet better performance than Tomeklinks, random undersampling, and edited nearest neighbors technique. Jain and his colleagues in Jain et al. (2017, 2020) applied genetic algorithms to improve the recognition rate of diseased patients while maintaining high correct prediction of healthy patients. Their undersampling-based evolutionary optimization reduced the majority class samples by maximizing the geometric mean, significantly improving the classification performance. Table 5 summarizes the main ideas of the oversampling techniques proposed in the reviewed literature and other information.

4.2.3 Hybrid methods and comparative studies of resampling techniques

Hybrid techniques are uncommonly used to deal with imbalanced medical data by combining undersampling the majority class and oversampling the minority class. Comparably, studies contrasted various sampling techniques to reduce class discrepancy.

Resampling boosted the accuracy of liver disease detection (Arbain and Balakrishnan 2019). Fahmi et al. applied random resampling after weighting samples using the class distribution's inverse proportions, which achieved superior performance than SMOTE (Fahmi et al. 2022). Hybridization of ROSE for majority and minority class and K-means to select boundary samples with SVM classifier improved the prediction of all diseases in Zhang and Chen (2019b).

SMOTE is commonly combined with various undersampling techniques (Shi et al. 2022; Xu et al. 2020; Wosiak and Karbowiak 2017). SMOTE-ENN with logistic regression remarkably identified chronic kidney patients at risk of end-stage and exceeded the Cox proportional hazard model (Shi et al. 2022). The authors in Xu et al. (2020) repeatedly

Table 5 Undersampling techniques

In-text citation	Key idea	Year
Rahman and Davis (2013)	A cluster-based undersampling method	2013
Parvin et al. (2013)	Balancing based on clustering	2013
Dai and Hua (2016)	Comparing different imbalanced machine learning methods	2016
Babar and Ade (2016)	An MLP based on undersampling	2016
Salman and Vomlel (2017)	Balancing based on clustering and instance weighting	2017
Neocleous et al. (2016)	Undersampling using k-means clustering	2017
Li et al. (2018)	Balancing by splitting the majority samples	2018
Richter and Khoshgoftaar (2018)	Random undersampling and feature selection	2018
Jain et al. (2017)	Optimized Evolutionary Under Sampling (OEUS)	2018
Zhang et al. (2020)	Balancing method by combining Gaussian Naïve Bayes as an estimator and entropy as a query	2020
Wang et al. (2020)	Undersampling based on Self-paced learning approach	2020
Vuttipittayamongkol and Elyan (2020b)	Undersampling based on recursive search neighboring	2020
Al-Shamaa et al. (2020)	A Hellinger distance-based undersampling method (HDUS)	2020
Jain et al. (2020)	Undersampling based on Genetic algorithms	2023
Kamaladevi and Venkatraman (2021)	Undersampling noise samples and borderline samples	2021
Iori et al. (2022)	Feature selection with undersampling the majority class	2022
Augustine and Jereesh (2022)	Clustering the Majority class and Random undersampling	2022
Ibrahim (2022)	Undersampling based on Clustering based undersampling using a Salp swarm optimization	2022

changed the oversampling ratio of SMOTE by the misclassification rate of trained RF on a subset of data and combined it with ENN. This hybrid method minimized the MCC (Matthews Correlation Coefficient) and statistically demonstrated significant performance compared to different data-level methods. The classification performance based on the Hybridisation of SMOTE with random undersampling fluctuated in Wosiak and Karbowiak (2017). However, SMOTE with Tomek Links showed superior performance (Zeng et al. 2016).

Few novel hybrid sampling methods were designed for imbalanced medical data (Babar 2021; Vuttipittayamongkol and Elyan 2020a). Babar and Ade (2016), the authors combined the MLPUS with the Majority Weighted Minority Oversampling Technique (MWMOTE), which assigns selection weights to important and borderline minority samples and then synthesizes new samples using clustering. A clustering approach was used further in the generation of synthetic samples. Investigation illustrates the better performance of the combination compared to MLPUS and MWMOTE separately. The authors in Vuttipittayamongkol and Elyan (2020a) eliminated the majority of instances based on the overlapping degree and oversampled minority instances in borderline regions using Borderline SMOTE; they attained high performance based on boosting,

Frequent studies compared sampling techniques in cancer diagnosis (Fotouhi et al. 2019), no-show cases detection (Krishnan and Sangar 2021), stroke diagnosis (Alamsyah et al. 2021), pediatric acute-conditions detection (Wilk et al. 2016), chronic kidney disease prediction (Yildirim 2017), heart disease prediction (Fernando et al. 2022), Lymph node metastasis prediction in stage T1 Lung adenocarcinoma (Lv et al. 2022), osteoporosis detection (Werner et al. 2016), predicting the risk of chronic kidney disease in cardiovascular disease patients (Vinothini and Baghavathi Priya 2020), and multi-minority medical data (Shilaskar and Ghatol 2019), however, results varied depending on the data used and experiment configurations.

The hybrid approach in imbalanced medical data seems to be less considered compared to advances in sampling techniques. Moreover, comparisons of sampling techniques yield to select the best, yet a balancing technique's outcome could vary based on many factors, including the medical data used. Table 6 describes the hybrid techniques in a nutshell.

5 Learning level

Modifications concerning the learning algorithms are grouped under this section and further classified into subgroups depending on the similarities in the used algorithm as described in the following.

5.1 Cost-sensitive learning

It attributes specific costs for misclassifying minority and majority samples. The misclassification costs are unknown; however, the cost matrix is usually inversely proportional to the distribution of classes in the original data. Therefore, more attention is given to the minority class.

Normally, cost-sensitive learning in medical data outperforms cost-insensitive learning (Wu et al. 2020; He et al. 2016; Phankokkruad 2020; Nguyen et al. 2020). Radial basis neural network (RBF-NN) based sample distribution adaptive cost function in Wu et al. (2020) exceeded different forms of RBF-NN, ensemble methods based on RBF, and single classifiers. He et al. used cost-sensitive neural networks and the cost as part of gradient descent

Table 6 Hybrid data level techniques

In-text citation	Key idea	Year
Tavares et al. (2013)	Comparing SMOTE, SVM for balancing (Farquad and Bose 2012), and a weighted SVM to assign error costs for each class (W.SVM)	2013
Werner et al. (2016)	Feature reduction with resampling techniques	2016
Wilk et al. (2016)	Comparing sampling techniques	2016
Yildirim (2017)	Comparing sampling techniques	2017
Wosiak and Karbowiak (2017)	SMOTE, random undersampling, and their combination	2017
Shilaskar and Ghatol (2019)	Comparing sampling techniques for multiclass imbalance	2019
Fotouhi et al. (2019)	Comparison of sampling techniques	2019
Zeng et al. (2016)	SMOTE with Tomek Links	2016
Arbain and Balakrishnan (2019)	Random sampling	2019
Zhang and Chen (2019b)	Hybridization of ROSE and sample selection by K-means	2019
Xu et al. (2020)	ENN and M-SMOTE	2020
Vinothini and Baghavathi Priya (2020)	Balancing by SMOTE and random undersampling separately	2020
Vuttipittayamongkol and Elyan (2020a)	Undersampling based on fuzzy clustering and borderline SMOTE	2020
Babar (2021)	Multiple Layer Perceptron Under-Sampling (MLPUS) and the Majority Weighted Minority Oversampling Technique	2021
Krishnan and Sangar (2021)	Comparison of sampling techniques	2021
Al-Shamaa et al. (2020)	Comparison of SMOTE and Nearmiss methods	2022
Fernando et al. (2022)	Comparing various sampling methods	2022
Lv et al. (2022)	Comparing two feature subsets and sampling techniques	2022
Shi et al. (2022)	SMOTE-ENN	2022
Fahmi et al. (2022)	Class weighting and balancing with Random resampling and SMOTE	2022

(He et al. 2016); investigation showed its minimal costs and significant accuracy. Cost-sensitive XGBoost model with the tuning of class weights effectively diagnosed breast cancer (Phankokkruad 2020). Likewise, a cost-sensitive version of Multiple Layer Perceptron and convolutional neural networks outperformed in detecting Inflammatory Bowel Disease (IBD) (Nguyen et al. 2020). However, some traditional ML algorithms yielded comparable results to developed cost-sensitive models, the decision rules algorithm and the ensemble of cost-sensitive SVM indistinguishably performed (Zięba 2014). While Decision Tree and Logistic regression achieved better accuracy than their corresponding cost-sensitive models (Mienye and Sun 2021).

Some research newly defined the cost matrix (Huo et al. 2022; Zhu et al. 2018; Belarouci et al. 2016; Wan et al. 2014). The authors in Belarouci et al. (2016) introduced a version of the least mean square algorithm to associate weights to different samples according to the errors, and investigations illustrated its superiority over SMOTE in breast cancer detection. Recently, Huo et al. used neural networks and set the misclassification costs as learnable parameters which released high performance in risk prediction in binary and multi-class classification (Huo et al. 2022). Class weights random forest based on class weighting for each classifier with threshold voting gave very optimistic results in Zhu et al.

(2018); while attributing weights based on a scoring function (RankCost) in Wan et al. (2014) outperformed cost-sensitive decision trees and Adaboost.

5.2 Optimization techniques

Recent methods applied Genetic algorithms to handle imbalanced medical data (Jain et al. 2020; Nalluri et al. 2020). Jain et al. (2020) optimized the specificity and sensitivity, where two models were constructed by employing the NSGA II algorithm and combined for the prediction of minority and majority samples. While the hybrid evolutionary learning with multiobjective exceeded optimization methods (Nalluri et al. 2020).

5.3 Simple classifier

It consists of using conventional machine learning algorithms to classify imbalanced medical data, which may include postprocessing or preprocessing procedures to tackle the imbalance issue and boost the classification performance.

Hyperparameter tuning with SVM models improved patient detection sometimes (Ksiaa et al. 2021), while performed similarly to cost-sensitive learning in Alzheimer's prediction (Zhang et al. 2022). Contrast classification strategy-based feature elimination demonstrated superior results compared to decision trees, and SVM (Dhanusha et al. 2022). Modification on the used classifiers released good results (Alves et al. 2023). Alves et al. developed a generalization of complementary loglog link functions for binary regression that better fitted the data than binomial models (Alves et al. 2023). Differently, Kumar and Thakur proposed A fuzzy learning approach hybridizing adaptive and neighbor-weighted KNN for liver disease detection that outpaced Fuzzy Adaptive KNN (Kumar and Thakur 2019).

5.4 Ensemble learning

This approach combines a set of single classifiers to perform classification tasks. There are three types of ensembles: bagging, boosting, and stacking. Bagging consists of building multiple single classifiers on different samples of the primary dataset and then combining their prediction with some basic statistics. Boosting is an iterative approach combining weak learners where each focuses on the misclassified instances by the previous one and generates predictions using a weighted average of constructed models. Finally, stacking is based on stacking different classifier types built on the same dataset and aggregating their predictions using another model (combiner).

Various ensemble learning classifiers effectively diagnosed the disease in imbalanced data (Zhao et al. 2022; Wei et al. 2017; Bhattacharya et al. 2017; Potharaju and Sreedevi 2016). Weighted ensemble-based Knn algorithm with feature extraction released remarkable results in identifying the stage of Parkinson's disease (Zhao et al. 2022). Similarly, ensemble Knn based with the relief-F method for feature selection accurately predicted the responses of breast cancer patients to neoadjuvant chemotherapy (Gao et al. 2018). Whereas the authors in Wei et al. (2017) used XGBoost based on EasyEnsemble, investigations demonstrated its high results in large-scale imbalanced diabetes data. Bhattacharya et al. (2017), the authors balanced the training subsets and employed a hierarchical Meta classification method, Experiments showed the high performance of random forest hierarchical meta-classifier in detecting later stages of chronic kidney disease that exceeded random oversampling and SMOTE. The majority voting ensemble of AdaBoost and Logistic regression outperformed AdaBoost and Logistic regression in heart disease detection (Rath et al. 2022). While ensemble by bootstrapping the majority class with a replacement and majority voting considerably detects different types of Parkinson's disease (Roy et al. 2023). In contrast, Zhao et al. (2021) ensembles various machine learning algorithms, where AdaBoost and XGBoost comparably outpaced other ensemble models. Mathew and Obradovic (2013) used homomorphic encryption to secure multi-party computation with a distribution voting ensemble if collected encrypted data was imbalanced, illustrating the superiority of ensemble models over baseline models.

Random forest revealed significant results compared to boosting and bagging techniques in the prediction of malaria disease (Sajana and Narasingarao 2018b) and thyroid (Çinaroğlu 2017). Differently, the authors in Guo et al. (2018) used an ensemble of rotation trees (ERT) including undersampling and feature extraction, and investigations showed, statistically, the excellent performance of ERT compared to EasyEnsemble, Random Undersampling Random Forest (RURF), BalanceCascade, and bagging. While in Potharaju and Sreedevi (2016) the authors developed ensembles of rule-based algorithms on SMOTE-balanced data, the experiments showed the optimal accuracy of AdaBoost.

5.5 Deep learning algorithms

Modification of the structure and parameters of neural networks and deep learning algorithms is found as an approach to tackle class imbalance in medical data and improve the classification performance (Ghorbani et al. 2022; Izonin et al. 2022; Liu et al. 2019; Sribhashyam et al. 2022). The authors in Ghorbani et al. (2022) combined a Graph convolutional network (GCN) algorithm with weighting networks and employed an iterative adversarial training process, demonstrating stability and superior performance compared to other GCN methods. An improved imbalanced probability neural network (IPNN) by Izonin et al. (2022) yielded high performance. Liu et al. (2019), the authors automated hyperparameter optimization (AutoHPO) of deep neural network (DNN) including dimensionality reduction using PCA K-means and majority instance selection with batch reweighting using online learning; investigation demonstrated the excellence of AutoHPO based on DNN compared to DNN, XGB, etc. ResNet and GRU with weighted focal loss function exceeded ResNet in multi-class heart disease detection (Rong et al. 2020). A stacked denoising autoencoder (SDA) for anomaly detection excelled LSTM, SVM, MLP with Borderline SMOTE, and SVM with SMOTE (Alhassan et al. 2018). Recently, Sribhashyam et al. used multi-instance neural network architecture that exceeded state-of-the-art methods for disease diagnosis (Sribhashyam et al. 2022).

5.6 Unsupervised learning

Unsupervised learning approaches showed high performance and interpretability; however, it is uncommonly used (Zhou and Wong 2021; Chan et al. 2017). Chan et al. (2017), the authors used a pattern discovery and heuristic optimization of the geometric mean, which significantly performed and bettered logistic regression. Lately, the authors in Zhou and Wong (2021) identified relevant patterns, for which they established a matrix representing the frequency of co-occurrence of pairs-values (like in association rules). Then, they build another matrix representing the frequency deviation from the default frequencies (the parallel of the covariance matrix in PCA). They decomposed this matrix into several

PCs and then projected these pairs of values in the sub-space. Then, they selected clusters (patterns). Experiments demonstrated the outperformance of the proposed algorithm over CART, Naive Bayes, and logistic regression.

Regarding structured medical data, deep learning is yet to be explored as a potential solution for class imbalance where many reasons may pop up, like the insufficiency of medical data or the model complexity. Another emerging research line is pattern recognition. A descriptive table (Table 7) provides all information about learning level techniques, like the year, the title, and the main idea.

6 Combined techniques and comparison of different approaches

Combining learning and data-level approaches is considered to treat imbalanced medical data. Studies contrasting different approaches or suggesting combined techniques are quite frequent as learning approaches in the last decade's literature.

Recently, studies combined deep learning approaches with sampling techniques and exceeded the state-of-the-art techniques (Feng and Li 2021; Woźniak et al. 2023). Feng and Li (2021), the authors optimized the borderline SMOTE and ADASYN combination αDBASMOTE where only minority samples in danger set are synthesized and used DenseNet convolutional neural network. Investigation illustrated the higher performance of αDBASMOTE over Borderline SMOTE and ADASYN. The authors in Woźniak et al. (2023) combined oversampling by ADASYN and SMOTE with undersampling by Tomek-Links and used a Bidirectional Long Short-Term Memory deep learning model which output promising results. Rath et al. ensembled LSTM and GAN based on GAN for data generation, and the investigations showed excellent results in heart disease detection (Rath et al. 2021). SVM based on the active learning approach relied on the degree of the instance's importance and yielded superior performance (Lee et al. 2015). Likewise, Suresh et al. (2022) used Radius SMOTE for balancing and Convolutional generative adversarial network for data generation with a modified CNN model, experimentation illustrates its optimal performance and lower computational time.

Preprocessing was integrated into class imbalance approaches (Cheng et al. 2022; Hallaji et al. 2021). Cheng et al. (2022) denoised signals and combined multi-scale features along with ADASYN for balancing different categories of Electrocardiogram (ECG). While Britto and Ali (2021) proposed balancing and augmenting the data and a deep learning model with adaptive weighting for minority classes. Hallaji et al. (2021) compared an adversarial imputation classification network (AICN) with hybrid models encompassing sampling with data imputation techniques. Miss-Forest was the most performant in imputation, and SMOTE was the best in balancing techniques, while AICN outperformed and showed stability in different missing value ratios. Ensemble learning combined with different approaches better-handled class imbalance in medical data (Gan et al. 2020; Gupta and Gupta 2022). AdaCost with tree-augmented naive Bayes network outpaced AdaCost variants (Gan et al. 2020), whereas experiments in Gupta and Gupta (2022) demonstrated the high performance of boosted ensemble stacking. Oversampling with Ensemble of PNN and weighted voting significantly outperformed PNN, biased random forest, and random undersampling boosting (Yuan et al. 2021). Liu et al. used hybrid sampling by SMOTE and Cross validated committee filter, then an ensemble of SVM with optimized weighted voting using simulated annealing genetic algorithm (SAGA) (Liu et al. 2020); investigation illustrated its optimal performance compared to the state-of-the-art classification models.

Table 7 Learning level techniques for imbalanced medical datasets

In-text citation	Approach	Key idea	Year
Wan et al. (2014)	Cost-sensitive	Cost-sensitive based on a scoring function (RankCost)	2014
Zięba (2014)	Cost-sensitive	Ensemble of cost-sensitive SVM classifiers	2014
Potharaju and Sreedevi (2016)	Ensemble	Ensemble learning methods with rule-based algorithms	2016
Belarouci et al. (2016)	Cost-sensitive	Samples weighting based on a version of the least mean square (LMS) algorithm	2016
Chan et al. (2017)	Unsupervised learning	Pattern discovery approach with a heuristic optimization method	2017
Chan et al. (2017)	Unsupervised learning	Pattern discovery approach with a heuristic optimization method	2017
Çinaroğlu (2017)	Ensemble	Comparing ensemble learning methods	2017
Mathew and Obradovic (2013)	Ensemble	A distribution voting ensemble model	2017
Wei et al. (2017)	Ensemble	An EasyEnsemble method based on XGBoost	2017
Bhattacharya et al. (2017)	Ensemble	A hierarchical Meta classification method	2017
He et al. (2016)	Cost-sensitive	A cost-sensitive neural network for handling imbalanced medical data	2017
Zhu et al. (2018)	Cost-sensitive	Class weighting with a threshold voting	2018
Sajana and Narasingarao (2018b)	Ensemble	Comparing ensemble learning techniques	2018
Guo et al. (2018)	Ensemble	An ensemble of rotation trees (ERT)	2018
Gao et al. (2018)	Ensemble	Feature selection with ensemble learning	2018
Alhassan et al. (2018)	Deep learning	An anomaly detection using a deep learning approach	2019
Liu et al. (2019)	Deep learning	PCA K-means with an automated hyperparameter optimization (AutoHPO) of deep neural network (DNN)	2019
Rong et al. (2020)	Deep learning	ResNet and GRU with weighted focal loss function	2020
Jain et al. (2020)	Optimization techniques	NSGA II algorithm to optimize sensitivity and specificity	2020
Nalluri et al. (2020)	Optimization techniques	Hybrid Evolutionary algorithm with Multiobjective	2020
Wu et al. (2020)	Cost-sensitive	Cost-sensitive radial basis neural network (RBNN) and optimization of its structure and its parameters	2020
Nguyen et al. (2020)	Cost-sensitive	A cost-sensitive version of Multiple Layer Perceptron and convolutional neural networks	2020
Phankokkruad (2020)	Cost-sensitive	A cost sensitive XGBoost model	2020
Mienye and Sun (2021)	Cost-sensitive	ML Cost-sensitive classifiers and insensitive classifiers	2021

Table 7 (continued)

In-text citation	Approach	Key idea	Year
Zhou and Wong (2021)	Unsupervised learning	A pattern recognition approach	2021
Zhao et al. (2021)	Ensemble	Ensemble learning with various machine learning algorithms	2021
Ksiaa et al. (2021)	Simple classifier	hyperparameters tuning of SVM and LASVM	2021
Zhang et al. (2022)	Simple classifier	Parallelized hyperparameter fine-tuning process to optimize the SVM model	2022
Dhanusha et al. (2022)	Simple classifier	Contrast classification approach based on feature elimination using mutual information	2022
Alves et al. (2023)	Simple classifier	A generalization of complementary loglog link in binomial regression models	2022
Kumar and Thakur (2019)	Simple classifier	Fuzzy learning approach hybridizing adaptive and neighbor weighted KNN	2019
Zhao et al. (2021)	Ensemble	Weighted ensemble based KNN algorithm	2022
Huo et al. (2022)	Cost-sensitive	Training framework based on recategorizing the data and misclassification costs	2022
Ghorbani et al. (2022)	Deep learning	Graph convolutional network algorithm with weighting networks and employing an iterative adversarial training process	2022
Izonin et al. (2022)	Deep learning	Modification of the probabilistic neural network	2022
Sribhashyam et al. (2022)	Deep learning	A multi-instance neural network architecture	2022
Rath et al. (2022)	Ensemble	LR, AdaBoost, and SVM	2022
Roy et al. (2023)	Ensemble	Generating balanced subsets using XGBoost with bootstrapping and majority voting	2023

Sampling with ensemble learning combined in different manners effectively handled class imbalance in disease diagnosis (Naghavi et al. 2019; Kinal and Woźniak 2020; Li et al. 2021; Lamari et al. 2021). ADASYN for oversampling and the cost-sensitive ensemble classifier constructed on SVM, KNN, and MLP conquered deep learning-based models in freezing of gait (FoG) prediction (Naghavi et al. 2019). Dynamic ensemble selection, in particular, DES-KNN coupled with SMOTE, significantly treated non-severely unbalanced data (Kinal and Woźniak 2020). Likewise, SMOTE-ENN sampling with dynamic classifier selection using META-DES exceeded the META-DES on imbalanced data (Lamari et al. 2021). Li et al. designed a harmonized-centred ensemble (HCE) approach that iteratively undersampled the majority class samples based on their classification hardness level (Li et al. 2021). Investigations demonstrated the outperformance of HCE over the Under-Bagging method, RUSBoost method, and self-paced ensemble learning framework (SPE). A SMOTE-based stacked ensemble with Bayesian optimization for hyperparameters tuning released excellent results in breast cancer diagnosis (Cai et al. 2018). The combination of SMOTE with SVM and AdaBoost surpassed stacking and voting strategies (Wang et al. 2020). Undersampling using different techniques with AdaBoost for learning and prediction attained optimal results (Shaw et al. 2021). Feature extraction, along with random undersampling and XGBoost, effectively predicted acute kidney injury in intensive care unit patients and outperformed random oversampling, random forest, AdaBoost, KNN, and Naïve Bayes (Wang et al. 2020). Similarly, Liu et al. (2014) used random undersampling to train SVM classifiers and validated them on data synthesized by SMOTE accordingly specific weights were attributed to SVMs; investigation illustrated the effectiveness of the SVM ensemble in cardiac complications of patients with chest pain in the emergency at the hospital.

Modifications on the random forest algorithm had considerable results (Meher et al. 2014; Lyra et al. 2019). Meher et al. (2014) developed a combined random forest where each random forest was trained on a balanced subset of data clustered from the original data. According to experiments, the combined random forest outperformed weighted and biased random forests. A “nested forest” was developed by Lyra et al. (2019) using feature selection and reduction with random undersampling to create balanced subsets for decision tree training, and the best forests were used for sepsis prediction. Fujiwara et al. (2020), the authors used boosting weights to select misclassified majority samples iteratively in the next CART classifier and oversampled the minority samples based on their distribution. Experiments demonstrated the superior performance of the approach in severely imbalanced medical data compared to random undersampling with boosting and SMOTE. In contrast, the scholars in Silveira et al. (2022) combined manual oversampling by a nephrologist and automated oversampling by SMOTE and its variants, where the decision tree achieved superior and stable performance in the early detection of chronic kidney disease.

The research compared class imbalance strategies in disease diagnosis (Drosou et al. 2014; Gupta et al. 2021; Wang et al. 2023) had different outcomes. In comparisons of resampling and cost-sensitive learning approaches (Drosou et al. 2014), while SVM is used for classification, the best performance was achieved by hybrid sampling (SMOTE and random undersampling) with SVM. The authors in Gupta et al. (2021) examined various class imbalance techniques where extensive experiments illustrated the outperformance of weighted XGBoost and stacking ensemble of weighted classifiers in breast cancer diagnosis. Additionally, feature selection, SMOTE, and cost-sensitive learning were employed with a variety of machine learning classifiers (Wang et al. 2023); however, three strategies achieved the best results in identifying patients with chronic obstructive pulmonary disease: cost-sensitive logistic regression, cost-sensitive SVM, and logistic regression with SMOTE.

Feature selection noticeably improved the classification performance in imbalanced medical data (Porwik et al. 2016; Špečkauskienė 2016; Lijun et al. 2018; Razzaghi et al. 2019). Wrappers for feature selection with parallel ensemble based on a weighted Knn achieved better and more stable accuracy than C4.5 and naïve Bayes in multi-class imbalanced and incomplete HCV data (Porwik et al. 2016). Feature selection outperformed Oversampling with SMOTE in multi-class Parkinson's disease detection (Špečkauskienė 2016) where the Clinical Decision Support system identified the best feature subset in Špečkauskienė (2011). Lijun et al. (2018) combined elastic net for feature selection and hybrid sampling using SMOTE and Random undersampling and used SVM multi-class investigations showed the superior overall accuracy achieved. Differently, ensemble learning methods with SMOTE and feature selection outperformed single classifiers particularly random forest and bagging yielded the highest results (Razzaghi et al. 2019). Tang et al. (2021), the authors combined feature selection and dimensionality reduction for biological data in breast cancer diagnosis and designed a twice-competitional ensemble method (TCEM) to select the optimal model, where results were promising. Cheng and Wang applied Particle Swarm Optimization (PSO) for feature selection with SMOTE and Random forest and achieved considerable breast cancer diagnosis results (Cheng and Wang 2020).

Optimization techniques were integrated into different approaches and largely improved the medical diagnosis (Shilaskar et al. 2017; Sadrawi et al. 2018; Desuky et al. 2021). Shilaskar et al. (2017) combined hybrid sampling with a modified particle swarm optimization to optimize the kernel function of SVM. The authors in (Sadrawi et al. 2018) used Fuzzy C-mean clustering to undersample the majority class and genetic algorithms to optimize the activation combination of the ensemble of activated ANN models. Including diversity within the ensemble and GA optimization yielded better results than single classifiers. Sampling using crossover genetic operator with adaptive boosting proposed by Desuky et al. (2021) improved classification performance better than SMOTE and safe level SMOTE (SLSMOTE). Feature selection and Principal Component Analysis with random oversampling and Ensemble voting exceeded SMOTE, SMOTE-ENN, and SMOTE-Tomek links (Alashban and Abubacker 2020). Srinivas et al. used rough set theory based on fuzzy c-mean clustering which exceeded the rough fuzzy classifier in heart disease detection (Srinivas et al. 2014). Table 8 is a descriptive table of all the combined techniques proposed for imbalanced medical data.

7 Synthesis of research outcomes on imbalanced medical datasets

Several benchmarking imbalanced datasets appear in the studied medical diagnosis research. Among the frequently medical diagnostics imbalanced data, we overview results on those frequently studied, namely: "Pima Diabetes Dataset", "Wisconsin Diagnostic Breast Cancer (WDBC)", "Wisconsin Prognostic Breast Cancer (WPBC)", "Haberman Dataset", "SPECT Heart Dataset", "Breast Cancer Dataset", "Indian Liver Patient Dataset (ILPD)", "Hepatitis-C Dataset", "Cervical Cancer Dataset", "Heart Disease Dataset", "Breast Cancer Wisconsin Original Dataset", "Parkinson's Disease Dataset", "New Thyroid Dataset", "Chronic Kidney Disease Dataset", "Thoracic Surgery Dataset", "Liver Disorder Dataset", "Mammographic Mass Dataset". This synthesis consolidates the findings

Table 8 Combined techniques for imbalanced medical data

In-text citation	Key idea	Year
Drosou et al. (2014)	Comparing resampling and cost-sensitive learning approaches	2014
Srinivas et al. (2014)	Fuzzy c-mean clustering and rough set theory for rules generation	2014
Meher et al. (2014)	A combined random forest-based clustering	2014
Liu et al. (2014)	Hybrid sampling with SVM ensemble	2014
Lee et al. (2015)	Pairing active learning and SVM in four different approaches	2015
Špečkauskienė (2016)	Feature selection with oversampling using SMOTE	2015
Porwik et al. (2016)	Feature selection methods with a modified version of Knn	2016
Shilaskar et al. (2017)	Oversampling using Euler's distance criterion and undersampling + SVM optimized using genetic algorithms	2017
Sadrawi et al. (2018)	Fuzzy C-mean clustering (FCM) + Ensemble + genetic algorithms for optimization	2018
Lijun et al. (2018)	Feature selection and oversampling	2018
Cai et al. (2018)	A hybridization of sampling with ensemble learning	2018
Razzaghi et al. (2019)	Ensemble learning, feature selection, and sampling	2019
Naghavi et al. (2019)	ADASYN with cost-sensitive learning (bagging and boosting)	2019
Lyra et al. (2019)	"Nested forest" based on feature selection and reduction and data augmentation	2019
Cheng and Wang (2020)	SMOTE, Particle Swarm Optimization PSO for attribute selection, and MetaCost sensitive learning	2020
Gan et al. (2020)	AdaCost with a tree augmented Naïve Bayes network	2020
Fujiwara et al. (2020)	Hybrid sampling and boosting	2020
Alashban and Abubacker (2020)	Sampling techniques and voting ensemble	2020
Kinal and Woźniak (2020)	Dynamic ensemble selection DES-KNN with SMOTE	2020
Liu et al. (2020)	SMOTE combined with Cross validated committee filter CVCF with SVM ensemble	2020
Wang et al. (2020)	Feature selection, SMOTE, cost-sensitive and ensemble learning	2020
Wang et al. (2020)	Ensemble learning based on feature extraction and random undersampling	2020
Gupta and Gupta (2022)	Sampling techniques with stacking basic classifiers and ensembles	2021
Hallaji et al. (2021)	Comparison of data imputation techniques and balancing techniques	2021
Li et al. (2021)	Undersampling with ensemble learning	2021

Table 8 (continued)

In-text citation	Key idea	Year
Lamari et al. (2021)	Hybrid Sampling SMOTE-ENN with dynamic classifier selection	2021
Tang et al. (2021)	A three-stage feature selection strategy (TSFS) with a twice competitional ensemble (TCEM) method	2021
Yuan et al. (2021)	KNN-based undersampling, SMOTE and ensembles	2021
Shaw et al. (2021)	Under-sampling the majority class with AdaBoost	2021
Desuky et al. (2021)	Sampling with ensemble based on Crossover genetic operator	2021
Feng and Li (2021)	An optimized version of borderline SMOTE and ADASYN αDBASMOTE	2021
Rath et al. (2022)	Deep learning and ensemble	2021
Gupta et al. (2021)	Examining multiple class imbalance techniques	2021
Cheng et al. (2022)	Balancing with ADASYN	2022
Britto and Ali (2021)	Data generation + data oversampling + data augmentation	2022
Silveira et al. (2022)	Ensemble and non-ensemble machine learning classifiers based on manual and automated data augmentation	2022
Suresh et al. (2022)	Radius SMOTE, deep convolutional generative adversarial network, and a modified CNN model	2023
Woźniak et al. (2023)	ADASYN, SMOTE, and undersampling using Tomek-Links	2023
Wang et al. (2023)	Feature selection, SMOTE, and cost-sensitive learning	2023

from research utilized key imbalanced medical datasets, providing a cohesive understanding of how these datasets are analyzed within the framework of class imbalance.

This analysis is contextual, relying on the employed class imbalance methodology by the research authors and its performance quantified in terms of evaluation metrics they selected. Those experimental details were the most explicitly reported across the literature; clarifications on the underlying methodological procedures could enhance the informativeness of observations. Thus, we attempt to bridge the theoretical frameworks of machine learning with their practical applications in medical diagnostics, using an observatory approach to offer a detailed overview of current practices and performance metrics, highlighting the utilization and effectiveness of these methods in different medical contexts without drawing new conclusions or conducting experimental analysis. It is important to note that this synthesis cannot be classified as experimental or deeply analytical due to several constraints. Consequently, our reflections on the synthesis setting up and context are mentioned accordingly.

Eleven research papers on medical diagnosis in imbalanced data have employed the "Breast Cancer Wisconsin Original Dataset" in experimentation. Table 9 summarizes the results of each research work and mentions the used approach in tackling the class imbalance issue. While this dataset presents an imbalance ratio of 1.90, various class imbalance methods have been used to tackle this imbalance. The learning approach is the most prevalent and yields excellent performance in classifying breast cancer, where combined techniques are the most implemented (Yuan et al. 2021; Kinal and Woźniak 2020; Suresh et al. 2022; Cai et al. 2018) compared to cost-sensitive methods (Wu et al. 2020), ensemble methods (Guo et al. 2018), and optimization techniques (Nalluri et al. 2020). Scholars have used data-level approaches, though less frequently than previous approaches, the outcomes are considerable performance in terms of different metrics where we found a feature-level method (Zhang and Chen 2019a), an oversampling method (Mustafa et al. 2017), an undersampling method (Vuttipittayamongkol and Elyan 2020b), hybrid method (Zhang and Chen 2019b). There are slight differences in performance metrics observed. However, the effectiveness of a method can be influenced by numerous factors, including the specific

Table 9 Results on "Breast Cancer Wisconsin Original Dataset"

Research	Approach	Acc	AUC	Sens	Spec	*F*-value	GM	Prec
Yuan et al. (2021)	Combined techniques				0.96			
Wu et al. (2020)	Cost-sensitive	0.96	0.99					
Zhang and Chen (2019a)	Feature level	0.99		1	0.99		0.99	
Kinal and Woźniak (2020)	Combined techniques	0.99		0.95		0.96	0.97	0.99
Mustafa et al. (2017)	Oversampling	0.94	0.89					
Vuttipittayamongkol and Elyan (2020b)	Undersampling			0.98	0.93	0.93	0.95	
Zhang and Chen (2019b)	Hybrid	0.97	0.98	1	0.96		0.98	
Guo et al. (2018)	Ensemble		0.97	0.99		0.95	0.97	
Suresh et al. (2022)	Combined techniques	0.94		0.93	0.93	0.96		0.95
Cai et al. (2018)	Combined techniques	0.98	0.98		0.97			
Nalluri et al. (2020)	Optimization techniques	1		1	1			

Acc. accuracy, *AUC* area under ROC curve, *Sens.* sensitivity, *Spec.* specificity, *GM* geometric mean, *Prec.* Precision characteristics of the data, the complexity of the model, and the research goals. In this analysis of the 'Breast Cancer Wisconsin Original Dataset,' we observe subtle variations in performance metrics among the different methodologies employed. Despite these variations, the overall classification performance remains considerable, demonstrating robustness in addressing class imbalances within this dataset.

Table 10 summarizes the findings from eleven distinct studies on the "Heart Disease Dataset," each employing different strategies to tackle the challenges of class imbalance in medical diagnostics. This dataset exhibits an imbalance ratio of 1.20; other versions of the datasets exist that could be differently imbalanced. The researchers experimenting always refer to the version presented in Table 1 unless other details are reported. This dataset has seen a variety of approaches, with combined techniques being particularly prevalent, as demonstrated in the works by Gan et al. (2020), Kinal and Woźniak (2020), Shilaskar et al. (2017), Desuky et al. (2021) and Srinivas et al. (2014), which display a range of outcomes across key metrics such as accuracy, sensitivity, specificity, and more. Other approaches include undersampling (Jain et al. 2020), which yielded high accuracy and sensitivity, and oversampling (Rodriguez-Almeida et al. 2022), although specific performance metrics for the latter are not reported; whereas optimization techniques employed by Nalluri et al. (2020) showed superior performance with nearly perfect metrics, indicating potential advantages depending on the specific methodological implementations and study goals. The hybrid approach by Shilaskar and Ghatol (2019) and optimization efforts by Chan et al. (2017) also added to the diversity of results, though with mixed effectiveness. This analysis reveals variations in how different methods perform under the constraints of the same dataset, reflecting a spectrum of effectiveness in the tools and strategies deployed. Despite these differences, the collective outcomes contribute significantly to advancing the diagnostic capabilities for heart disease, illustrating the value of diverse methodological approaches in enhancing overall classification performance.

Table 11 synthesizes the outcomes from five research studies on the "Cervical Cancer Dataset," focusing on various methodologies used for cervical cancer diagnosis. This dataset, in particular, has the highest class imbalance among reference medical datasets, as seen in Table 1. It is observed a predominant reliance on combined techniques, as employed by Gan et al. (2020), Gupta and Gupta (2022), Kinal and Woźniak (2020), and Woźniak et al. (2023). Each study shows differing levels of effectiveness across metrics such as accuracy,

Table 10 Results on "Heart Disease Dataset"

Research	Approach	Acc	Sens	Spec	GM	AUC	*F*-value	Prec
Jain et al. (2020)	Undersampling	0.92	1	0.89	0.94			
Gan et al. (2020)	Combined techniques	0.80				0.88		
Kinal and Woźniak (2020)	Combined techniques	0.91	0.93		0.90		0.92	0.92
Rodriguez-Almeida et al. (2022)	Oversampling	–	–	–	–	–	–	–
Shilaskar and Ghatol (2019)	Hybrid	0.77						0.83
Chan et al. (2017)	Optimization techniques		0.55		0.58		0.25	0.16
Shilaskar et al. (2017)	Combined techniques	0.83	0.83	0.96		0.83	0.83	0.84
Desuky et al. (2021)	Combined techniques	0.70	0.86		0.70		0.72	0.61
Srinivas et al. (2014)	Combined techniques	0.81	0.62	1				
Nalluri et al. (2020)	Optimization techniques	0.99	0.98	0.99				

Table 11 Results on "Cervical Cancer Dataset"

Research	Approach	Acc	AUC	Prec	Sens	*F*-value	GM	Spec
Mienye and Sun (2021)	Cost-sensitive	0.98		1	1	1		
Gan et al. (2020)	Combined techniques	0.92	0.87					
Gupta and Gupta (2022)	Combined techniques	0.98						
Kinal and Woźniak (2020)	Combined techniques	0.92		0.67	0.68	0.65	0.72	
Woźniak et al. (2023)	Combined techniques	0.99		0.99	0.93	0.97		1

AUC, precision, sensitivity, *F*-value, geometric mean, and specificity. Mienye and Sun (2021) utilized a cost-sensitive approach, which stands out with exceptional results—achieving perfect scores in accuracy, AUC, precision, and sensitivity. In contrast, the combined techniques exhibit a range of performances, with Woźniak et al. (2023) demonstrating notably high efficacy, almost reaching optimal scores across all evaluated metrics. This array of studies reflects the effectiveness of different learning strategies in diagnosing cervical cancer. It highlights the diversity in methodological success and underlines the particular strengths of more nuanced approaches, like the cost-sensitive method showcased by Mienye and Sun. Overall, two main learning methods are observed, whereas the aggregated findings from these studies highlight their contribution to advancements in cervical cancer diagnostics concerning the studied data.

Table 12 assembles findings from multiple research studies that have applied various approaches to the "Hepatitis Dataset," characterized by an imbalance ratio of 3.84. This summary highlights how the twelve research papers employed different methods to address the challenges inherent in the imbalanced data, employing ensemble, cost-sensitive, hybrid, undersampling, oversampling, feature-level, combined techniques, and optimization strategies. Among the methodologies, the feature-level approach by Polat (2018) stands out with perfect scores across all metrics, showcasing the potential of finely tuned feature engineering in such contexts. Similarly, optimization techniques used by Nalluri et al. (2020) and

Table 12 Results on "Hepatitis Dataset"

Research	Approach	Acc	Sens	Spec	GM	*F*-value	AUC	Prec
Guo et al. (2018)	Ensemble		0.81		0.78	0.62	0.80	
Wan et al. (2014)	Cost-sensitive		0.84			0.62		0.5
Wosiak and Karbowiak (2017)	Hybrid	0.73	0.71					
Babar and Ade (2016)	Undersampling	0.97	0.86		0.93			0.98
Naseriparsa et al. (2020)	Oversampling		0.84			0.88	0.94	0.93
Polat (2018)	Feature level	1	1			1	1	1
Kamaladevi and Venkatraman (2021)	Undersampling	0.83	0.83			0.97	0.95	
Babar (2021)	Hybrid	0.98						
Jain et al. (2020)	Undersampling	0.97	0.91	1	0.94			
Gupta and Gupta (2022)	Combined techniques	0.99						
Yuan et al. (2021)	Combined techniques				0.73			
Nalluri et al. (2020)	Optimization techniques	0.99	1	0.98				

combined techniques by Gupta and Gupta (2022) demonstrated high effectiveness, with near-perfect accuracy and other metrics. Conversely, approaches like the ensemble by Guo et al. (2018) and the hybrid technique by Wosiak and Karbowiak (2017) yielded more modest results, accentuating the variability in the efficacy of different methodologies within the same imbalanced dataset. The undersampling methods, particularly those implemented by Babar and Ade (2016) and Jain et al. (2020), showed remarkable improvements in handling class imbalance, reflected in their high accuracy and specificity. This aggregation of studies illustrates a broad expanse of success in managing class imbalance of the dataset, with some methods showing considerable effectiveness while others highlight areas for potential improvement.

Table 13 gathers the performance metrics from several studies that utilized the "Indian Liver Patient Dataset (ILPD)" to address its class imbalance of 2.49. The table provides a broad overview of the effectiveness of different class imbalance approaches, including simple classifiers, undersampling, combined techniques, and optimization strategies. The results demonstrate a range of effectiveness across methodologies. Combined Techniques employed by Gan et al. (2020), Yuan et al. (2021), and Kinal and Woźniak (2020), these methods yielded mixed results. Gan et al. and Yuan et al. reported relatively lower specificities and sensitivities, while Kinal and Woźniak achieved a high specificity of 0.95, indicating that the success of combined techniques can vary significantly based on their specific configurations and the aspects of the data they prioritize. On the other hand, the simple classifier approach by Kumar and Thakur (2019) showed a high *F*-value and precision, suggesting that even straightforward models can perform effectively within this dataset. Undersampling, proposed by Jain et al. (2017, 2020), showed improvements in specificity and sensitivity, indicating its utility in enhancing model accuracy by addressing data imbalance. Meanwhile, Nalluri et al. (2020) applied optimization techniques, which resulted in balanced performance across all metrics. This table of findings across different studies illuminates the varied effectiveness of each methodology in handling the dataset's imbalance. Each demonstrates high values in some metrics and lower values in others. It illustrates the necessity of selecting an appropriate method based on specific dataset characteristics and desired outcomes in diagnostic accuracy.

Table 14 assembles the results from diverse research methodologies to diagnose breast cancer using the "Breast Cancer Dataset." This dataset's imbalance of 2.38 has prompted researchers to employ mixed techniques, including undersampling, cost-sensitive methods, ensemble approaches, hybrid strategies, and combined techniques. Undersampling is mostly used with varied results, as illustrated by Al-Shamaa et al. (2020) with modest

Table 13 Results on "Indian Liver Patient Dataset (ILPD)"

Research	Approach	Spec	Sens	Acc	*F*-value	Prec	AUC	GM
Kumar and Thakur (2019)	Simple classifier	0.65	0.88	0.84	0.90	0.91		
Jain et al. (2017)	Undersampling							0.85
Jain et al. (2020)	Undersampling	0.75	0.94	0.8				0.84
Gan et al. (2020)	Combined techniques			0.68			0.68	
Yuan et al. (2021)	Combined techniques							0.59
Kinal and Woźniak (2020)	Combined techniques		0.95	0.76	0.82	0.72		0.62
Nalluri et al. (2020)	Optimization techniques	0.77	0.84	0.79				

Table 14 Results on “Breast Cancer Dataset”

Research	Approach	Spec	Sens	Acc	Prec	F-value	AUC	GM
Al-Shamaa et al. (2020)	Undersampling	0.66	0.77		0.43	0.56		
Babar and Ade (2016)	Undersampling		0.97	0.96	0.96			0.96
Wan et al. (2014)	Cost-sensitive		0.49		0.49	0.49		
Zięba (2014)	Cost-sensitive	0.70	0.55	0.66				0.62
Guo et al. (2018)	Ensemble		0.59			0.50	0.64	0.62
Ibrahim (2022)	Undersampling	0.93	0.93		0.94	0.95	0.96	
Babar (2021)	Hybrid			0.90				
Jain et al. (2020)	Undersampling	0.68	0.92	0.75				0.79
Yuan et al. (2021)	Combined techniques							0.68

outcomes in specificity and sensitivity, contrasting significantly with Ibrahim (2022), which achieved high values across these metrics. Similarly, Babar and Ade (2016) and Jain et al. (2020) also utilized undersampling, resulting in a particularly strong performance from the former. Wan et al. 2014 and Zięba (2014) applied cost-sensitive methods, showing lower performance metrics. Guo et al. (2018) employed an ensemble approach, yielding middling results, which suggest a complexity in achieving higher predictive accuracy through this method. In other studies, specific performance metrics are not fully detailed, highlighting a need for more comprehensive results. Babar (2021) implemented a hybrid method, achieving considerable accuracy, and Yuan et al. (2021) explored combined techniques and achieved an average trade-off of sensitivity and specificity. Significant variability in the literature outcomes is observed, suggesting the ongoing challenges and complexities in diagnosing breast cancer in this particular imbalanced dataset.

Table 15 showcases the results from seven distinct studies that have applied various methodologies to the “SPECT Heart Dataset,” which has an imbalance ratio of 3.85. These methodologies encompass miscellaneous methods to improve diagnostic accuracy and address the dataset’s imbalance. The study by Polat (2018) indicates the efficacy of feature level adjustments, yielding excellent performance metrics. Jain et al. (2017, 2020) both employed undersampling techniques. While the later study provides specific details on performance metrics like specificity, sensitivity, and accuracy—all marked consistently at 0.88—Jain et al. (2017) attained a geometric mean of 0.91, suggesting effective handling of class imbalances. Babar (2021) utilized a hybrid approach and achieved an accuracy

Table 15 Results on “SPECT Heart Dataset”

Research	Approach	Spec	Sens	Acc	Prec	F-value	AUC	GM
Polat (2018)	Feature level		0.96	0.96	0.97	0.96	0.99	
Jain et al. (2017)	Undersampling							0.91
Babar (2021)	Hybrid			0.84				
Jain et al. (2020)	Undersampling	0.88	0.88	0.88				0.88
Liu et al. (2020)	Combined techniques		0.88		0.90	0.89	0.79	0.75
Kinal and Woźniak (2020)	Combined techniques		0.68	0.79	0.84	0.74		0.76
Nalluri et al. (2020)	Optimization techniques	0.93	0.97	0.95				

of 0.84. Liu et al. (2020) and Kinal and Woźniak (2020) both opted for combined techniques, with varying levels of success across specific and general performance metrics. Nalluri et al. (2020) implemented optimization techniques, resulting in impressive specificity, sensitivity, and accuracy scores. The synthesis in Table 15 reflects the diverse strategies researchers can employ to tackle diagnostic challenges and underscores the complexity of achieving high accuracy in class imbalances.

Table 16 groups the results of research studies exploring various techniques to address the challenges presented by the "Haberman Dataset," which exhibits an imbalance ratio of 2.78. This imbalance influences the choice of methodological approaches, including sampling strategies, learning techniques, and combined techniques. The outcomes of sampling methods vary, while the oversampling method in Xu et al. (2021) effectively mitigates class disparity, achieving optimal results in sensitivity and specificity, the results of Wang et al. (2013) denote a modest value of sensitivity, and the undersampling technique proposed in Jain et al. (2020) indicate relatively considerable performance. Other studies report their results in one metric, Jain et al. (2017) proposing an undersampling reported a high precision value, and Xu et al. (2020) used hybrid sampling reflected in a high *F*-value, suggesting an effective balance between recall and precision. Mienye and Sun (2021) adopts a cost-sensitive technique, achieving notable sensitivity and precision. Leveraged by Ghorbani et al. (2022) and Izonin et al. (2022), deep learning models excel in discerning complex patterns, with Izonin's findings excelling in sensitivity and precision. Liu et al. (2020) and Desuky et al. (2021) employ combined techniques, achieving balanced values across various metrics. Nalluri et al. (2020) explores optimization techniques for class imbalance, leading to average metrics values. This synthesis stresses diverse approaches to enhancing model accuracy against the Haberman Dataset's imbalance. We observe better performance in terms of sensitivity along recent studies achieved and significant differences between the findings of the literature on this dataset, while few achieved excellent performance, others potentially need to tackle effectively class imbalance in particular and understanding of the medical data in general.

The reviewed medical diagnosis research results in imbalanced data employing the WPBC dataset are presented in Table 17. knowing that this dataset exhibits an imbalance of 3.21, we observe that five studies proposed sampling methods to handle the class imbalance in the data, where the outcomes of the research proposing oversampling (Xu et al.

Table 16 Results on "Haberman Dataset"

Research	Approach	Sens	Spec	*F*-value	Acc	GM	Prec	AUC
Wang et al. (2013)	Oversampling	0.57	0.85		0.73	0.85		
Jain et al. (2017)	Undersampling						0.85	
Xu et al. (2021)	Oversampling	1	1					
Xu et al. (2020)	Hybrid			0.97				
Mienye and Sun (2021)	Cost-sensitive	0.9		0.88	0.80		0.87	
Jain et al. (2020)	Undersampling	0.8	0.8		0.79	0.79		
Ghorbani et al. (2022)	Deep learning				0.75			0.61
Izonin et al. (2022)	Deep learning	0.90		0.90			0.91	
Liu et al. (2020)	Combined techniques	0.84		0.86		0.80	0.84	0.84
Desuky et al. (2021)	Combined techniques	0.63		0.74	0.67	0.71	0.89	
Nalluri et al. (2020)	Optimization techniques	0.71	0.75		0.79			

Table 17 Results on "Wisconsin Prognostic Breast Cancer (WPBC)"

Research	Approach	Sens	Spec	*F*-value	Prec	GM	AUC	Acc
Jain et al. (2017)	Undersampling					0.96		
Zhang and Chen (2019b)	Hybrid	0.69	0.58			0.63	0.64	0.64
Xu et al. (2021)	Oversampling	1	1					
Xu et al. (2020)	Hybrid			0.98				
Jain et al. (2020)	Undersampling	1	0.93			0.96		0.94
Yuan et al. (2021)	Combined techniques					0.69		
Liu et al. (2020)	Combined techniques	0.77		0.66	0.72	0.88	0.75	
Nalluri et al. (2020)	Optimization techniques	0.96	0.96					0.96

2021) indicate optimal performance, other studies employing undersampling (Jain et al. 2017, 2020) and hybrid (Xu et al. 2020) imply significant values of performance metrics, whereas the study (Zhang and Chen 2019b) implementing hybrid sampling indicate modest performance. The findings of research works (Yuan et al. 2021; Liu et al. 2020) proposing combined techniques appear modest, although Liu et al.'s approach indicates a better balance between sensitivity and specificity. However, the effectiveness of Nalluri et al. (2020) that implemented optimization technique for class imbalance is superior regarding reported metrics. Yet, the analysis of the noted results in handling the imbalance in the WPBC dataset points to the presence of diverse class imbalance methods and the variation in the performance of implemented methodologies.

Table 18 presents the findings from research studies utilizing the "Wisconsin Diagnostics Breast Cancer (WDBC)" dataset. These studies have implemented a variety of approaches, including combined techniques, algorithmic-level modifications, and preprocessing methods, to address the challenges associated with this dataset's imbalance of 1.68 and improve diagnostic accuracy. Combining techniques dominate the research landscape and are used in eight studies, including (Shaw et al. 2021), which achieved perfect scores across sensitivity, accuracy, and precision metrics. Similarly, Kinal and Woźniak (2020), Desuky et al. (2021), Cai et al. (2018) and Liu et al. (2020) showed excellent outcomes, underscoring the efficacy of these approaches in optimizing performance across several metrics. However, Yuan et al. (2021), Cheng and Wang (2020) and Gupta and Gupta (2022) indicated high outcomes in terms of specific performance metrics. Four studies employed cost-sensitive methods to enhance model sensitivity to cost discrepancies between classes. Belarouci et al. (2016) achieved ideal results in all evaluated metrics, illustrating the potential of these methods to balance predictive accuracy and cost considerations. Zhu et al. (2018), Wu et al. (2020) and Phankokkruad (2020) also showed significant improvements, particularly in specificity and *F*-values; while Nalluri et al. (2020) implemented optimization techniques and showed impressive results. Several studies utilized oversampling to correct imbalances in dataset representation (Naseriparsa et al. 2020; Luo et al. 2021; Lan et al. 2022; Xu et al. 2021), with Xu et al. achieving near-perfect sensitivity and specificity. Similarly, studies implemented hybrid sampling (Zhang and Chen 2019b; Xu et al. 2020) and undersampling (Jain et al. 2020), though presented superior performance, Xu et al. (2020) perfectly score across multiple metrics. On the other hand, Zhang and Chen (2019a) applied feature-level modifications, resulting in high marks across sensitivity, specificity, and accuracy. Differently, Izonin et al. (2022) implemented deep learning to deal with the class imbalance and achieved remarkable results. The diverse methodologies listed in Table 18 reflect diverse strategies researchers employ to tackle the class imbalance issue of

Table 18 Results on "Wisconsin Diagnostics Breast Cancer (WDBC)"

Research	Approach	Sens	Spec	Acc	*F*-value	GM	AUC	Prec
Naseriparsa et al. (2020)	Oversampling	0.93			0.95		0.99	0.96
Shaw et al. (2021)	Combined techniques	1		1	1		0.99	1
Desuky et al. (2021)	Combined techniques	0.98		0.97	0.97	0.98	1	
Cheng and Wang (2020)	Combined techniques			0.95				
Cai et al. (2018)	Combined techniques		0.98	0.97			0.97	
Phankokkruad (2020)	Cost-sensitive			0.99			0.99	
Belarouci et al. (2016)	Cost-sensitive	1	1	1		1		
Zhu et al. (2018)	Cost-sensitive	0.98			0.9		0.92	
Zhang and Chen (2019b)	Hybrid	1	1	1		1	1	
Zhang and Chen (2019a)	Feature level	0.96	0.98	0.98		0.97		
Xu et al. (2021)	Oversampling	1	0.99					
Luo et al. (2021)	Oversampling				0.97			
Lan et al. (2022)	Oversampling			0.94				
Xu et al. (2020)	Hybrid				0.99			
Wu et al. (2020)	Cost-sensitive			0.95			0.99	
Jain et al. (2020)	Undersampling	1	0.96	0.97		0.97		
Izonin et al. (2022)	Deep learning	0.97			0.97			0.97
Gupta and Gupta (2022)	Combined techniques			0.99				
Yuan et al. (2021)	Combined techniques					0.96		
Liu et al. (2020)	Combined techniques	0.97			0.96	0.95	0.97	0.94
Kinal and Woźniak (2020)	Combined techniques	0.98		0.98	0.99	0.98		0.99
Nalluri et al. (2020)	Optimization techniques	1	1	1				

the WDBC dataset. Although combined techniques show particular prevalence, different approaches suggested optimal effectiveness. This overview not only underlines the variability in method effectiveness but also highlights the ongoing advancements in breast cancer diagnostics, emphasizing the achievement in diagnostic accuracy.

Twenty-nine research articles have utilized the Pima Diabetes Dataset in experimentation. Table 19 summarizes all the experimental results of one feature-level method, fourteen sampling methods, nine algorithmic-level approaches, and six combined techniques. Table 19 summarizes the diverse research methodologies applied to the "Pima Diabetes Dataset," exhibiting an imbalance with a ratio of 1.87. The dataset's imbalance and relevance have urged the adoption of various approaches to improve predictive accuracy and handle data imbalances: one feature-level method, fifteen sampling methods, seven learning-level approaches, and six combined techniques. The results of the oversampling method proposed by Rodriguez-Almeida et al. (2022) are inexplicitly mentioned. Combined techniques output an average geometric mean (Yuan et al. 2021). The two sampling techniques proposed by Zeng et al. (2016) and Hassan and Amiri (2019) and the two learning methods suggested by Ghorbani et al. (2022) and Wu et al. (2020) have a good overall performance in diabetes diagnosis with AUC values in the range (0.8–0.88). The three sampling methods (Xu et al. 2020), Babar (2021) and Mustafa et al. (2017) yield excellent global performance by achieving values greater than 0.98 in *F*-value, accuracy, and AUC, respectively. We categorize the remaining

Table 19 Results on “Pima Diabetes Dataset”

Research	Approach	Sens	Spec	*F*-value	GM	Acc	AUC	Prec
Hassan and Amiri (2019)	Oversampling						0.84	
Desuky et al. (2021)	Combined techniques	0.87		0.67	0.73	0.71		0.55
Guo et al. (2018)	Ensemble	0.78		0.68	0.75		0.75	
Suresh et al. (2022)	Combined techniques	0.92	0.93	0.93		0.94		0.93
Wan et al. (2014)	Cost-sensitive	0.80		0.69				0.60
Zeng et al. (2016)	Hybrid			0.80	0.80	0.80	0.88	
Zhang and Chen (2019b)	Hybrid	0.82	0.74		0.78	0.77	0.78	
Al-Shamaa et al. (2020)	Undersampling	0.91	0.66	0.70				0.56
Babar and Ade (2016)	Undersampling	0.85			0.86	0.86		0.86
Rodriguez-Almeida et al. (2022)	Oversampling	–	–	–	–	–	–	–
Naseriparsa et al. (2020)	Oversampling	0.79		0.8			0.87	0.82
Wang et al. (2013)	Oversampling	0.63	0.80		0.71	0.75		
Mustafa et al. (2017)	Oversampling					0.91	0.99	
Polat (2018)	Feature level	0.96		0.96		0.96		0.96
Xu et al. (2021)	Oversampling	0.99	0.98					
Kamaladevi and Venkatraman (2021)	Undersampling	0.78		0.97		0.75	0.94	
Ibrahim (2022)	Undersampling	0.78	0.78	0.78			0.82	0.79
Babar (2021)	Hybrid					0.98		
Xu et al. (2020)	Hybrid			0.99				
Mienye and Sun (2021)	Cost-sensitive	0.84		0.80		0.75		0.77
Wu et al. (2020)	Cost-sensitive					0.74	0.80	
Jain et al. (2020)	Undersampling	0.98	0.63		0.78	0.75		
Ghorbani et al. (2022)	Deep learning					0.74	0.8	
Izonin et al. (2022)	Deep learning	0.77		0.77				0.76
Yuan et al. (2021)	Combined techniques				0.78			
Liu et al. (2020)	Combined techniques	0.72		0.65	0.75		0.75	0.64
Kinal and Woźniak (2020)	Combined techniques	0.84		0.87	0.87	0.89		0.91
Lamari et al. (2021)	Combined techniques	0.77	0.68	0.76	0.71		0.79	
Nalluri et al. (2020)	Optimization techniques	0.90	0.88			0.89		

works in the literature into four groups based on their sensitivity score. The oversampling method proposed by Wang et al. (2013) poorly recognizes diabetes patients. Further, The methodologies proposed by Guo et al. (2018), Liu et al. (2020), Naseriparsa et al. (2020), Kamaladevi and Venkatraman (2021), Lamari et al. (2021), Ibrahim (2022) and Izonin et al. (2022) averagely identify patients with diabetes. Approaches in Wan et al. (2014), Babar and Ade (2016), Zhang and Chen (2019b), Kinal and Woźniak (2020), Nalluri et al. (2020), Desuky et al. (2021) and Mienye and Sun (2021) attain a considerable detection of patients with the disease. Finally, the following methods excellently classify the target group (diseased): optimization technique in Jain et al. (2020), feature level in Polat (2018), undersampling by Al-Shamaa et al. (2020), combined techniques by Suresh et al. (2022), and oversampling in Xu et al. (2021). Nevertheless, the latter shows excellent specificity as well. The findings outlined in Table 19

Table 20 Results on “Parkinson’s Disease Dataset”

Research	Approach	Sens	Spec	*F*-value	Acc	GM	Prec	AUC
Polat (2018)	Feature level	1		1	1		1	1
Sug (2016)	Oversampling	0.99	0.91		0.96	0.95		
Jain et al. (2017)	Undersampling					0.95		
Zeng et al. (2016)	Hybrid			0.93	0.93	0.93		0.99
Jain et al. (2020)	Undersampling	1	1		1	1		
Lamari et al. (2021)	Combined techniques	0.74	0.62	0.73		0.66		0.74
Nalluri et al. (2020)	Optimization techniques	0.99	1		0.99			

Table 21 Results on “New Thyroid Dataset”

Research	Approach	Sens	Spec	*F*-value	Acc	GM	Prec	AUC
Guo et al. (2018)	Ensemble	1		1	1		1	0.99
Shaw et al. (2021)	Combined techniques	0.99		0.94		0.98		0.98
Liu et al. (2020)	Combined techniques	1		1		1	1	1
Kinal and Woźniak (2020)	Combined techniques	1	1	0.97	0.99	1	0.95	
Shilaskar et al. (2017)	Combined techniques	1	1	1	1	1	1	1
Nalluri et al. (2020)	Optimization techniques	1	0.99		0.99			

reveal the class imbalance strategies designed for diabetes prediction using the Pima Diabetes Dataset. The varied methodologies underscore the dynamic nature of medical diagnostics research, where each approach provides distinct advantages and faces specific challenges. This synthesis recaps the diverse strategies employed and highlights the expanding field as researchers seek more accurate and efficient diagnostic models.

Table 20 exposes the results from several studies that have utilized the “Parkinson’s Disease Dataset” to evaluate different diagnostic approaches. Knowing that this dataset exhibits an imbalance of 3.06, these studies encompass a range of methodologies: five preprocessing approaches, one optimization technique, and one combined techniques approach have been proposed to handle the class imbalance in the disease data. We notice the inferior performance of the combined techniques strategy in diseased and non-diseased patients’ detection. Moreover, the three sampling methods suggested by Sug (2016), Zeng et al. (2016) and Jain et al. (2017) and the optimization techniques by Nalluri et al. (2020) achieve an excellent tradeoff between diagnosing both patients with/without Parkinson’s ($0.93 < Sens < 0.99$). Furthermore, the feature level method in Polat (2018) and the undersampling method in Jain et al. (2020) correctly identify all cases. This analysis indicates the diverse methodologies implemented and the variation in their effectiveness in classifying Parkinson’s cases, where optimal performance is unveiled by some methods while other methods struggle to show comparable performance.

Table 21 presents the literature findings, within the review time range, realized in thyroid diagnosis using the “New Thyroid Dataset”: four combined techniques approaches and two learning ones. All the methods significantly diagnose patients (Sensitivity > 0.99). However, the following combined techniques (Shilaskar et al. 2017; Liu et al. 2020) optimally perform according to sensitivity, specificity, accuracy, and geometric mean.

Table 22 Results on "Chronic Kidney Disease Dataset"

Research	Approach	Sens	Spec	*F*-value	Acc	GM	Prec
Rodriguez-Almeida et al. (2022)	Oversampling	–	–	–	–	–	–
Yildirim (2017)	Hybrid	0.99		0.99			0.99
Mienye and Sun (2021)	Cost-sensitive	1		0.99	0.98		0.99
Jain et al. (2020)	Undersampling	1	1		1	1	
Suresh et al. (2022)	Combined techniques	0.93	0.78	0.93	0.94		0.92

Table 23 Results on "Thoracic Surgery Dataset"

Research	Approach	Sens	Spec	*F*-value	Acc	GM	Prec	AUC
Kinal and Woźniak (2020)	Combined techniques	0.62		0.68	0.83	0.74	0.86	
Chan et al. (2017)	Optimization techniques	0.57		0.3		0.58	0.21	
Jain et al. (2020)	Undersampling	0.76	0.92		0.9	0.83		
Zięba (2014)	Cost-Sensitive	0.35	0.73		0.66	0.51		
Jain et al. (2017)	Undersampling					0.86		
Al-Shamaa et al. (2020)	Undersampling	0.80	0.4	0.39			0.25	
Vuttipittayamongkol and Elyan (2020b)	Undersampling	0.95	0.05	0.25		0.23		
Polat (2018)	Feature level	1		1	1		1	1
Nalluri et al. (2020)	Optimization techniques	0.87	0.90		0.89			

Effectiveness in handling the class imbalance among the various proposed methodologies is observed, indicating overcoming the challenges related to the mentioned dataset.

Among the reviewed literature, five studies analyzed the "Chronic Kidney Disease Dataset"; their outcomes are shown in Table 22. The results of the oversampling method proposed by Rodriguez-Almeida et al. (2022) are unclearly mentioned. On the other hand, significant performance has been reached by the hybrid method proposed by Yildirim (2017) and the combined techniques proposed by Suresh et al. (2022). Both the undersampling method proposed by Jain et al. (2020) and the learning approach suggested by Mienye and Sun (2021) perfectly diagnose patients with chronic kidney disease; however, the former optimally identifies non-diseased patients. Various methodologies adopted different class imbalance methods; however, broad significant performance is observed in experimenting with this chronic kidney disease data.

Table 23 summarizes the results of the proposed approaches experimenting with the "Thoracic Surgery Dataset". The dataset presents an imbalance of 5.14; therefore, studies proposed different class imbalance methods within their classification methodologies. The optimization technique proposed by Chan et al. (2017) obtains low values of both sensitivity and specificity; significant detection of diseased patients associated with a poor detection of non-diseased patients or the opposite is noticed in the following three studies: undersampling techniques proposed in Al-Shamaa et al. (2020) and Vuttipittayamongkol and Elyan (2020b), the optimization technique in Nalluri et al. (2020), and the cost-sensitive method in Zięba (2014). The combined techniques have released average geometric mean value (Kinal and Woźniak 2020), while relatively superior values have been resulted

Table 24 Results on "Liver Disorder Dataset"

Research	Approach	Sens	Spec	*F*-value	Acc	GM	Prec	AUC
Wang et al. (2013)	Oversampling	0.58	0.78		0.7	0.67		
Babar and Ade (2016)	Undersampling	0.84			0.84	0.84	0.83	
Babar (2021)	Hybrid				0.96			
Wu et al. (2020)	Cost-sensitive				0.68			0.72
Shaw et al. (2021)	Combined techniques	0.85		0.86	0.86		0.86	0.86
Nalluri et al. (2020)	Optimization techniques	0.82	0.82		0.82			

Table 25 Results on "Mammographic Mass Dataset"

Research	Approach	Sens	*F*-value	Acc	GM	Prec	AUC
Babar and Ade (2016)	Undersampling	0.86	0.84	0.85	0.85	0.85	
Babar (2021)	Hybrid			0.88			
Zhu et al. (2018)	Cost-sensitive	0.86	0.76				0.77
Desuky et al. (2021)	Combined techniques	0.88	0.78	0.77	0.77	0.70	

in by undersampling methods (Jain et al. 2017, 2020). Finally, optimal performance has been attained by the feature-level method (Polat 2018). Regarding the "Thoracic Surgery Dataset", differences in the effectiveness of outlined methodologies are observed globally throughout the analysis; while the approach of class imbalance is noted, other jointly affecting factors exist in the context.

Regarding the investigation in Liver Disorder detection, six research works have been conducted using the "Liver Disorder Dataset", and Table 24 shows their outcomes. The hybrid method proposed by Babar (2021) has the highest accuracy value demonstrating its superior global performance. On the other hand, the accuracy and the area under curve values of the cost-sensitive approach proposed by Wu et al. (2020) refer to its average overall performance. However, in medical diagnosis models, particularly with imbalanced data, more specific metrics, like the sensitivity and geometric mean, are considered in performance assessment. Thus we notice the inferior performance of the oversampling method by Wang et al. (2013) in diseased patients diagnosis with a sensitivity of (0.58). Moreover, the undersampling technique, the optimization technique, and the combined techniques approach (Babar and Ade 2016; Nalluri et al. 2020; Shaw et al. 2021) outcome good values of sensitivity (> 0.82); however, the latter has higher values of accuracy, precision, and AUC and a better sensitivity which may be attributed to its significant performance in identifying patients with/out liver disorder. Various class imbalance methods were proposed, nonetheless, we notice that overall classification performance on this Liver Disorder data could be further improved.

Four distinct studies experimented with the "Mammographic Mass Dataset"; their findings are in Table 25. The hybrid strategy in Babar (2021) has attained the best accuracy of (0.88), unveiling its overall good performance. A good ratio of lesion detection is achieved by the undersampling method in Babar and Ade (2016), the cost-sensitive method in Zhu et al. (2018), and the combined techniques (Desuky et al. 2021), while the two first have equal sensitivity which refers diagnosing the malignant breast cancer lesion. The undersampling method has a good compromise between sensitivity and specificity, with a higher geometric mean and accuracy. Although few studies utilized Mammographic Mass Data, we observe the relatively considerable performance of the proposed methodologies globally.

8 Discussion

Of greater interest is exploring observations made through contextual analysis in this section. Thus, we discuss reflections on the synthesis of the outcomes of previous research on the reference medical datasets to point out speculative insights on methodological concerns and practical aspects in investigating class imbalance in medical data.

Methodologies performance considering the class imbalance methods For each medical dataset, we selected approaches that showed high performance; thus, twenty-two highly-performing methods on seventeen datasets, meaning various research works outcome similar optimal results in some medical datasets. The research by Polat (2018), proposing a feature level method, indicated optimal performance in handling class imbalance in three imbalanced medical datasets, namely: “Hepatitis-C Dataset”, “Parkinson’s Disease Dataset”, and “Thoracic Surgery Dataset”; where the data points of each attribute are weighted using similarity and clustering considering the class label.

Similarly, In breast cancer diagnosis using both the “Breast Cancer Wisconsin Original Dataset” and the “Wisconsin Diagnostic Breast Cancer (WDBC)” and in heart disease detection using the “SPECT Heart Dataset”, the research based on optimization techniques proposed by Nalluri et al. showed the most effectiveness in classification. Briefly, the method of Nalluri et al. (2020) uses a hybrid EA with Multiobjective, the fitness function is SVM, along with two Multiobjective scenarios and population with non-dominated solutions and limit solutions. The oversampling method proposed by Xu et al. (2021) appeared to be the most successful approach in treating three imbalanced medical datasets, which are: “Haberman Dataset”, “Wisconsin Prognostic Breast Cancer (WPBC)”, and “Pima Dataset”. In detail, this method uses a filtered k-means clustering to identify a new data matrix, which utilizes newly calculated sampling ratios and SMOTE to balance the data classes. The research adopting hybrid methods implied superior results in one dataset, “Wisconsin Diagnostic Breast Cancer (WDBC)”; this method hybridizes oversampling by ROSE and Sample selection by K-means to handle the imbalance in medical data (Zhang and Chen 2019b).

Overall undersampling techniques showed high classification performance in five medical datasets. The research proposing an undersampling method (Jain et al. 2020) based on Genetic algorithms could be perceived as the most efficient strategy for addressing the class imbalance in the following datasets: “Parkinson’s Disease Dataset”, “Chronic Kidney Disease Dataset”, and “Indian Liver Patient Dataset (ILPD)”; other studies proposed undersampling methods for class imbalance (Babar and Ade 2016; Vuttipittayamongkol and Elyan 2020b) respectively in “Breast Cancer Dataset” and “Heart Disease Dataset” outcomed the most promising results. The former is multiple-layer perceptron-based undersampling. At the same time, the latter Identifies the overlapping space of instances using recursive search neighbouring, then discards the majority instances in it to improve the visibility of minority instances. In cervical cancer diagnosis using the “Cervical Cancer Dataset”, the cost-sensitive approach suggested by Mienye & Sun, a cost-sensitive random forest classifier, indicated the optimal results. Whereas, in breast cancer diagnosis using the "Wisconsin Diagnostic Breast Cancer (WDBC)" dataset, the cost-sensitive method by Belarouci et al. (2016) suggested the most effectiveness as hybrid and combined techniques. It consists of a version of the least mean square (LMS) algorithm that associates weights to different samples according to the errors.

The approaches proposed in Shilaskar et al. (2017) and Liu et al. (2020) appear to be the most effective in thyroid detection using the "New Thyroid Dataset". Liu et al. (2020) proposed a SMOTE combined with a cross-validated committee filter (CVCF) and SVM ensemble, and Shilaskar et al. (2017) combined oversampling and undersampling along with SVM optimized using genetic algorithms. Moreover, the study, suggesting a combined techniques approach, by Shaw et al. (2021) outcomes excellent results along with that based on optimization techniques (Nalluri et al. 2020). Knowing that Shaw et al. under-sample the majority class with three different techniques and then combine the picked samples with the minority class with AdaBoost for prediction. Additionally, The research studies based on combined techniques (Shaw et al. 2021) and Desuky et al. (2021) likely surpass other approaches in two datasets: "Liver Disorder Dataset" and "Mammographic Mass Dataset", respectively, and releasing optimal diagnoses. The latter is Sampling with an ensemble based on a Crossover genetic operator to handle class imbalance.

Among the studies suggesting high classification performance in the medical reference dataset, the prevalence of the preprocessing-level methods theoretically owing to their extent of use in the reviewed literature, around sixty-one papers addressed the class imbalance proposing preprocessing, where hybrid sampling presented in 20 research works, undersampling in 18, and oversampling in 17. Besides, even the studies based on combined techniques, likely outperforming, utilize sampling techniques. Moreover, the research proposing feature-level methods indicates promising results, which could be a prominent research line, especially in sensitive clinical applications, by avoiding reliability issues of synthetic samples. On the other hand, Learning level methods are equally mentioned in research works reportedly efficient in some medical datasets. The distinct specifics of the datasets detailed in Table 1, coupled with the diversity of methodologies explored in the existing literature, suggest that the findings are context-dependent and may not be broadly applicable, emphasizing the need for cautious interpretation and an understanding of the limited scope.

Objectives in class imbalance research for medical applications Reference datasets presented in Table 1 are repetitively used for various methodological frameworks, whether for evaluating the class imbalance approach designed for diagnosing or studying a specific disease. A shared objective for those studies is the evaluation of the proposed approach over medical data exhibiting a certain degree of imbalance; while the objectives normally set in ML for medical diagnosis research are conditional to the given data and the medical application and relevance through the studied research the interchangeability between we observed a lack of specificity in how terms like 'diagnosis,' 'prediction,' 'classification,' and 'early detection' are employed interchangeably. This could be attributed to the overarching challenges of class imbalance, which seem to outweigh the need for clear differentiation in study objectives. Regardless of the stated goal, the primary concern often remains with the performance metrics of the learning algorithms due to the class imbalance, leading to a uniform approach in evaluating methodologies across different medical objectives. This issue is compounded by the general absence of transparent reporting in the literature, where distinctions between medical applications are often vague. Notably, this is less the case in works specifically targeting mortality prediction, which tend to demonstrate a clearer connection between methodological choices and their clinical implications. To enhance the clarity and applicability of research in this field, there is a need for more precise definitions of study objectives, specialized methodologies that directly address these objectives, and transparent reporting that links specific methodological approaches to their clinical outcomes.

Transparency in class imbalance approaches The literature often lacks detailed descriptions of datasets, methodologies, and experimental implementations, which limits the depth of analysis to an exploratory level. For instance, data-level methods such as sampling ratios frequently omit details like post-balancing data distribution. Even when aspects like evaluation techniques, preprocessing steps, and underlying learning algorithms are well-documented, they add layers of complexity that complicate straightforward observational synthesis. As such, including diverse details from the reviewed works increases the synthesis process's complexity and necessitates a more intensive investigative approach that transcends traditional observational efforts. This demands methodologies that delve beyond mere describing, requiring a rigorous examination of methodologies, results, and their interrelations within the broader research landscape to achieve a more comprehensive analysis.

Standardization issues in class imbalance The variability in class imbalance degrees across datasets reviewed herein spotlights a significant challenge in medical research. What may be deemed highly imbalanced in one study might only present as moderate in another, reflecting the quantitative differences and the diverse challenges each dataset presents. For example, slight imbalances in one dataset could be more problematic than severe imbalances in another, depending on factors such as the complexity of the medical conditions involved or data quality issues. This variability highlights the necessity for context-specific approaches in handling class imbalances, where the unique characteristics of each dataset are considered in the development and application of methodologies.

Furthermore, the absence of a universally accepted standard for quantifying the severity of class imbalance complicates the comparison of results across different studies and hampers the development of potentially broadly effective solutions. This lack of standardization calls for establishing clear metrics that could guide researchers in accurately classifying and reporting the degree of imbalance. Enhanced reporting standards and systematic analysis approaches are essential to facilitate a more consistent evaluation of method effectiveness across varied research contexts. By advocating for standardized quantification and comprehensive reporting, the research community can better understand the impact of class imbalance on medical diagnostics and develop more adaptable methodologies to improve the reliability and generalizability of outcomes in medical research.

9 Value and limitations

This comprehensive review of the literature addressing the issue of class imbalance involves the new detailed classification of class imbalance methods and informative statistics on the evaluation metrics and medical datasets and is further enhanced with practical insights by synthesizing the literature findings on the reference medical datasets with class imbalance. We aimed to extend the deep literature review with an overview of the experimental outcomes of proposed class imbalance methodologies that could not be reproducible; further, we intended to provide the reader with a contextual analysis describing the findings considering the found settings knowing that it was difficult to mention all the factors implemented in previous research due to general descriptions missing necessary configurations and methodological procedures.

Therefore, such a review in its experimental insights referring to the presented synthesis of research outcomes exhibits some limitations that could not be resolved in the current work due to correspondence issues with the principal question and the challenges it takes to establish a thorough comparative analysis controlling all the factors affecting the environment of experimenting with imbalanced medical data, to mention, but not limited to, the data size, the data dimensionality, the preprocessing procedures, the underlying learning algorithm, the imbalance ratio, the class imbalance method itself if involving other parameters such as the matrix of costs definition in cost-sensitive learning methods or the imbalance ratio in data level approaches.

Thus, the discussion of the overviewed findings is indicative and descriptive and states the need for an exhaustive experimental review to derive decisive and generalizable conclusions. Despite these limitations, our work maps out the landscape of existing research and emphasizes the variability and complexity of approaches, suggesting a compulsory need for standardization in research reporting and methodology. By highlighting these areas, we contribute to a deeper understanding of how class imbalance affects medical dataset analysis and point towards areas where further research and more refined methodologies are needed.

10 Trends and research directions

This section scrutinizes the predominant trends and emergent strategies in addressing class imbalance within medical datasets as identified in our comprehensive review of the past decade's literature. We feature key methodological innovations and the evolving paradigms that have shaped current approaches to managing imbalances in medical diagnostic data. Our analysis outlines the methodologies and links them to their potential impacts on enhancing diagnostic accuracy and clinical applicability.

Oversampling Researchers usually divide the minority class into three clusters: outliers (also called noise), safe samples, and in-danger or overlapped samples; when the distribution of each of the majority class and the minority class are overlapped, this consists of the borderline samples, known as in danger samples or overlapped samples. However safe cluster contains only samples that are in the minority distribution. Outliers or noisy samples are samples on the extreme side of the distribution, far from the mean distribution of minority samples. After this partition, some researchers only keep safe samples, oversample them Xu et al. (2021), and consider in-danger and outliers as noisy samples (deleted). However, in-danger samples or samples on the borderline are important in discriminating the minority from the majority, especially in our context, medical diagnosis, or medical applications in general. Other researchers (Han et al. 2019) adopt another partition of samples into four categories: noise samples, border samples, unstable samples, and stable samples, where only noisy samples are deleted. There is no unification in the partition of samples, which differentiates one sampling algorithm from another. Furthermore, as much as it depends on the sample's distribution, this partition needs to be explored in future research so that a partition is derived based on the data distribution automatically to retain the characteristics of the primary data.

Undersampling Research works like the Tversky similarity-based undersampling (Kamaladevi and Venkatraman 2021) and others remove the noise from the majority and the border samples, a major mistake. Knowing that samples at the borderline space also called the danger space, are hard to classify correctly, they are the most critical samples.

If the classifier learns to classify those samples, it will significantly succeed in classifying any new sample. This is because those samples contain the recognition patterns of both minority and majority samples. Nevertheless, they are still hard to learn from because it is where both distributions of minority class and majority class intersect and nearly exist. For that, samples in the border space can be exploited to improve the classification of imbalanced data rather than deleting them. So proper methods can be developed to address this issue.

Algorithmic solutions complexity with preprocessing simplicity (deep learning and ensemble) Another existing trend is to use ensemble learning or complex algorithms like deep learning algorithms (neural networks, graph neural networks) combined with optimization processes, without the preprocessing phase of minority samples. Such research works deal with class imbalance problems at the algorithmic level by optimizing the classification algorithm's parameters or/and structure rather than treating the imbalance at the data level. Also, similar works combine deep learning algorithms with cost-sensitive learning by adding misclassification weights to the training phase. As a result, the main common thing in this approach is using simple preprocessing techniques and focusing on the learning phase. However, the learning phase appears complex in several works like stacking, ensemble, deep learning algorithms with/without optimization, and cost-sensitive learning combined with deep learning or previously mentioned methods.

Genetic algorithms for optimization Another prevailing trend is the use of genetic algorithms in optimizing the learning classifier or the sampling technique. Even though researchers proposed multi-objective functions that are not well explored, which may be a future research direction, GA was used in undersampling and yielded good results; however, the proposed GA-based algorithms miss the optimization of parameters setting, which can significantly improve the performance of such methods.

SMOTE performance SMOTE is always used as a reference in comparative analysis in any work proposing a developed method. Reviewing all these results shows that SMOTE maintains stability and good performance patterns no matter how the class imbalance severity changes or the learning process is designed. Moreover, even if the newly proposed methods (sampling or algorithmic techniques) surpass SMOTE in some classification metrics, SMOTE still indicates better or similar results based on the remaining metrics. Nevertheless, a sampling method that exceeds SMOTE according to all classification assessment metrics is undiscovered, although the disadvantages of SMOTE techniques include synthesizing noisy and overlapped samples.

Feature selection Another approach in the literature chooses feature selection to tackle class imbalance in medical data and prove good results. However, this approach is not well explored as only some efforts of researchers in imbalanced medical datasets combine some feature selection techniques with improved classifiers. In our context, many reviewed papers include feature selection in the pre-processing phase, but how feature selection can be a performant solution in addressing class imbalance is a question that should be thoroughly discussed.

The compromise between sensitivity and specificity Another point regarding sampling in general and dealing with the imbalance in medical diagnosis is the problem of finding a compromise between correctly predicted diseased people and correctly detected non-diseased people, namely between sensitivity and specificity. The trade-off between those measures in our context is discussed by only one research; however, it is a long-lasting issue that is ignored. Future research may consider developing well-performing methods in classifying diseased and non-diseased people as an advanced level of improving existing approaches in imbalanced data classification. The reason for such a situation is that the unhealthy class represents rare cases. The focus is more on predicting unhealthy patients to provide early treatment and lessen the dangerous complications, so it is considered the class of interest. Nevertheless, intelligent systems of medical diagnosis or aid-medical diagnosis should be more careful towards both classes as an advanced level of intelligence.

Enhancing ensemble learning Whether modifying the ensemble selection like dynamic ensemble selection, modifying the structures of ensemble members, or making it cost-sensitive needs more investigation to evaluate its effectiveness; besides, stacking is sparingly found in the literature (Gupta and Gupta 2022). However, it shows considerable performance besides combining ensemble with cost-sensitive learning.

Simple classifiers Postprocessing (hyperparameters fine-tuning) or preprocessing procedures like feature selection show significant performance. Another research (Zhao et al. 2022) proposed a simple learning approach, an ensemble of KNN with weighting voting, that also leads to good results. Thus, simple, easy-to-implement and interpret, and unsophisticated algorithms without classic solutions for handling class imbalance resulted in significant accuracy and recall, as seen in simple classifiers reviewed papers.

Synthetic data and original data The use of synthetic data is prevalent in addressing class imbalance in medical datasets. However, ensuring that these data accurately reflect the real-world characteristics of original datasets is essential to prevent the introduction of biases that could compromise the fairness of medical diagnostics. Statistical tests to verify the similarity between synthetic and real data are necessary to maintain the integrity of medical models as initiated in Rodriguez-Almeida et al. (2022). This coherence is vital for the accuracy of the models and for ensuring that they do not perpetuate or exacerbate existing disparities in diagnosis outcomes. Future research should focus on developing methods that ensure both representative and equitable synthetic data, promoting fairness in medical diagnostics by adhering to rigorous standards that prevent bias and enhance the generalizability of research findings across diverse patient populations. This approach will support the broader goal of equitable healthcare by ensuring that advancements in medical diagnostics are accessible and beneficial to all population segments, thus upholding ethical standards in medical research and practice.

Interpretability and explainability In this last decade, many machine/deep learning algorithms have emerged to tackle the issue of class imbalance in medical diagnosis. We have observed a progressive evolution towards increasingly sophisticated and intricate models throughout the literature. While these algorithms frequently exhibit promising results in research environments, a significant disparity exists in their practical implementation within clinical settings. This discrepancy primarily stems from the need for more interpretability and trust among practitioners, especially in critical medical contexts. In light of these considerations, future research endeavours should prioritize the development of algorithms equipped to address imbalanced diagnoses while offering interpretability. Such models promise to enhance transparency in decision-making processes, thereby enabling greater understanding and trust among practitioners. This, in turn, paves the way for improved acceptance and adoption rates. Diverse approaches can be explored to achieve explainability, including employing model-agnostic techniques or incorporating post-hoc explanations. Such strategies facilitate domain experts' comprehension of complex model behaviours, even in cases where the proposed models lack inherent interpretability.

Computational efficiency and clinical deployment Another practical challenge associated with complex models in addressing imbalanced diagnoses is their computational efficiency, which directly influences their usability in clinical settings. By prioritizing computational efficiency in model development, researchers can effectively bridge the gap between sophisticated machine/deep learning models and their practical deployment by practitioners. This emphasis ensures that the models offer advanced capabilities and are feasible for real-world implementation.

Federated learning Another crucial research direction is addressing ethical concerns using federated learning models. This decentralized approach enables training models locally on distributed servers while safeguarding data privacy. Moreover, it proves advantageous in mitigating bias in data collection by training models across various geographic healthcare institutions. This broader representation holds the potential to yield more balanced models, particularly beneficial when class imbalance issues stem from bias in data collection rather than inherent population characteristics.

Deep learning approach in tabular imbalanced medical data It is exciting research lately, whether with data generation using GANs and their variants, graph-based deep learning approaches, or probabilistic neural networks that are newly suggested. Recent advancements in addressing class imbalance in medical data have seen researchers proposing sophisticated methodologies, necessitating a comparative analysis with traditional approaches to elucidate their differences better. Among these innovations, applying deep learning techniques such as Generative Adversarial Networks (GANs) for data generation—combined with classical machine learning algorithms, sampling, and cost-sensitive techniques—has yielded remarkable results.

Despite these technological advancements, the foundational aspect of data integrity remains critical. We cannot overstate the importance of establishing structured data collection designs that preserve the inherent population characteristics and ensure the representativeness of the collected sample. Such rigor in data collection is essential to avoid the injection of bias, which can skew the outcomes of even the most advanced analytical methods. As the field progresses, both cutting-edge technology and accurate data management must be harmonized to address the complexities of class imbalance in medical data fully. Additionally, integrating domain expertise in model training is crucial in ensuring these technologies are technically advanced and clinically relevant. Combining deep medical insights with innovative machine learning techniques enhances diagnostic tools' accuracy and applicability, supporting the ultimate goal of improving healthcare outcomes through more sophisticated and informed data science practices.

11 Mispractices and consensus on handling imbalanced data

This literature review revealed mispractices in class imbalance proposed strategies, particularly in medical data. These mispractices prevent the accurate evaluation of proposed methodologies and degrade any comparative study. In this section, we present the common fatal mistakes existing in literature and scholars still adopting in treating imbalanced data and propose the best practices instead. Besides, such best practices must be considered in this research line. There is an unveiled consensus amongst researchers on them. Thus, stating this consensus in our literature review is indispensable to advance the state of the art and yet, in the future, possess better and more effective tools to combat the class imbalance. Without treating these misconducts, any proposed methods may be inappropriately evaluated, yet future research will dismiss the starting points and falsely build on wrongly presented methods.

Overall performance measures in class imbalance methods Using general metrics to evaluate the performance of models in imbalanced data remains a critical issue. According to this literature review, multiple research works selected few metrics yet single metrics like accuracy. Relying only on accuracy, AUC-ROC, and F-measure uncover the real effectiveness of the model due to the imbalance in the used data. As a result, metrics reveal the model's performance in each class in the data. Therefore, a tendency to use sensitivity, specificity, and other metrics is required.

Data partition with data augmentation Augmenting the minority samples in an imbalanced dataset is a way to balance it. It is commonly used in literature and could be performed using any oversampling method. Usually, the selected sampling technique is applied to the training dataset. Hence, the machine learning model learns on balanced data where existing classes are equally represented. Conventional machine learning models are constructed on equally distributed data and expect the same in training datasets. Consequently, by sampling the training data, the learning algorithm gives equal attention to majority and minority classes. While data partition divides the data into training and test to select the best-performing model, only the training set participates in learning the model; the test set should be preprocessed like the training set. However, it should retain the data distribution characteristics as in real-world data. Testing the trained model on balanced data in the context of class imbalance leads to unrealistic results and misinformation on prediction performance. Additionally, proposing a new sampling method and using it before the train-test split will incorrectly tell us of its effectiveness. Thus, even a comparison with other research works is useless. Instead, highlighting the best practice when selecting oversampling to handle class imbalance appears necessary to prevent misconduct in future research.

Consensus on performance evaluation metrics in class imbalance Researchers in class imbalance should circulate the used evaluation metrics for future research purposes. Setting an ensemble of have-to-use metrics in treating imbalanced data appears unignorable. The set of metrics should involve a variety of metrics to measure the real performance of proposed approaches efficiently. As an attempt, we suggest the following: Sensitivity, Specificity, and Accuracy.

12 Conclusion

This paper presents the inaugural comprehensive review of the literature addressing the class imbalance in medical data, analyzing a decade's worth of research. Through a rigorous search methodology, 137 research articles were deemed relevant and subjected to a critical evaluation within a structured framework. Initially, the review introduces a novel classification of class imbalance methods, categorizing them into three primary approaches: preprocessing, learning, and combined techniques. This categorization facilitates a subtle exploration of contemporary techniques by further subdividing them into detailed subclasses.

Specifically, the learning approach is divided into six subclasses: cost-sensitive learning, optimization techniques, simple classifiers, ensemble learning, deep learning algorithms, and unsupervised learning. Similarly, the preprocessing approach comprises two detailed subclasses. The third category consists of combined techniques and comparative studies of different approaches. Furthermore, the paper provides an extensive overview and descriptive statistics of the medical datasets and evaluation metrics utilized in the reviewed literature, thoroughly examining current research practices and conventions.

Moreover, by synthesizing the outcomes of previous studies on reference medical datasets, this review provides an exploratory overview of the field's current state, identifying key trends and gaps that future research must address while clarifying related implications and the limited scope of our observatory reflections. The trends found in the literature have been comprehensively explained, and the prominent future research directions are pointed out, providing plausible research initiation points. Finally, we presented methodological strategies and procedural guidelines that can be implemented to ameliorate research studies in class imbalance, intending to augment the robustness, reliability, and generalizability of findings. The consensus should be broadly acknowledged to align communal measures toward devising optimal strategies to address the class imbalance issue.

Author Contributions Mabrouka Salmi: Conceptualization-Equal, Data duration-Lead, Investigation-Lead, Methodology-Equal, Visualization-Lead, Writing—original draft-Lead, Writing—review & editing-Equal. Dalia Atif: Methodology-Equal, Writing—review & editing-Equal. Ajith Abraham: Methodology-Equal, Writing—review & editing-Equal. Diego Oliva: Methodology-Equal, Writing—review & editing-Equal. Sebastian Ventura: Conceptualization-Equal, Funding acquisition-Lead, Supervision-Lead, Writing—review & editing-Equal.

Funding Funding for open access publishing: Universidad de Córdoba/CBUA. Spanish Ministry of Science and Innovation and the European Fund for Region Development, Grant: PID2020-115832-I00

Availability of data and materials Not applicable.

Code Availability Not applicable.

Declarations

Conflict of interest The authors declare no competing financial and/or non-financial interests about the described work.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Abd Elrahman SM, Abraham A (2013) A review of class imbalance problem. J Netw Innov Comput 1(2013):332–340

Alamsyah ARB, Anisa SR, Belinda NS, Setiawan A (2021) Smote and nearmiss methods for disease classification with unbalanced data: case study: Ifls 5. Proc Int Confer Data Sci Offic Stat 2021:305–314

Alashban M, Abubacker NF (2020) Blood glucose classification to identify a dietary plan for high-risk patients of coronary heart disease using imbalanced data techniques. In: Computational science and technology: 6th ICCST 2019, Kota Kinabalu, Malaysia, 29–30 August 2019. Springer, pp 445–455

Albuquerque J, Medeiros AM, Alves AC, Bourbon M, Antunes M (2022) Comparative study on the performance of different classification algorithms, combined with pre-and post-processing techniques to handle imbalanced data, in the diagnosis of adult patients with familial hypercholesterolemia. PLoS One 17(6):1–19

Alhassan Z, Budgen D, Alshammari R, Daghstani T, McGough AS, Al Moubayed N (2018) Stacked denoising autoencoders for mortality risk prediction using imbalanced clinical data. In: 2018 17th IEEE international conference on machine learning and applications (ICMLA). IEEE, pp 541–546

Ali H, Salleh MNM, Saedudin R, Hussain K, Mushtaq MF (2019) Imbalance class problems in data mining: a review. Indones J Electr Eng Comput Sci 14(3):1560–1571

Al-Shamaa ZZ, Kurnaz S, Duru AD, Peppa N, Mirnezami AH, Hamady ZZ et al (2020) The use of Hellinger distance undersampling model to improve the classification of disease class in imbalanced medical datasets. Appl Bion biomech 2020:1–10

Alves JS, Bazán JL, Arellano-Valle RB (2023) Flexible cloglog links for binomial regression models as an alternative for imbalanced medical data. Biom J 65(3):2100325

Arbain AN, Balakrishnan BYP (2019) A comparison of data mining algorithms for liver disease prediction on imbalanced data. Int J Data Sci Adv Analyt (ISSN 2563-4429) 1(1):1–11

Augustine J, Jereesh A (2022) An ensemble feature selection framework for the early non-invasive prediction of Parkinson's disease from imbalanced microarray data. In: Advances in computing and data sciences: 6th international conference, ICACDS 2022, Kurnool, India, April 22–23, 2022, revised selected papers, Part II. Springer, pp 1–11

Awon VK, Balloccu S, Wu Z, Reiter E, Helaouie R, Reforgiato Recupero D, Riboni D (2022) Data augmentation for reliability and fairness in counselling quality classification. In: Proceedings of the 1st workshop on scarce data in artificial intelligence for healthcare (SDAIH 2022). SciTePress

Babar V (2021) Classification of imbalanced data of medical diagnosis using sampling techniques. Commun Appl Electr 7:7–12

Babar V, Ade R (2016) A novel approach for handling imbalanced data in medical diagnosis using undersampling technique. Commun Appl Electron 5:36–42

Baniasadi A, Rezaeirad S, Zare H, Ghassemi MM (2020) Two-step imputation and adaboost-based classification for early prediction of sepsis on imbalanced clinical data. Crit Care Med 49(1):e91–e97

Belarouci S, Bouchikhi S, Chikh MA (2016) Comparative study of balancing methods: case of imbalanced medical data. Int J Biomed Eng Technol 21(3):247–263

Bhattacharya M, Jurkovitz C, Shatkay H (2017) Assessing chronic kidney disease from office visit records using hierarchical meta-classification of an imbalanced dataset. In: 2017 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, pp 663–670

Bi W, Ma R (2021) Unbalanced data set processing method for colorectal cancer prediction in tcm diagnosis. In: 2020 IEEE international conference on E-health networking, application & services (HEALTHCOM). IEEE, pp 1–6

Britto CF, Ali ARH (2021) Prostate cancer diagnosis model with the handling of multi-class imbalance through the adaptive weighting based deep learning model. EFFLATOUNIA-Multidiscipl J 5(2):3204–3212

Cai T, He H, Zhang W (2018) Breast cancer diagnosis using imbalanced learning and ensemble method. Appl Comput Math 7(3):146–154

Chan TM, Li Y, Chiau CC, Zhu J, Jiang J, Huo Y (2017) Imbalanced target prediction with pattern discovery on clinical data repositories. BMC Med Inform Decis Mak 17:1–12

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

Cheng CH, Wang YC (2020) A novel multi-combined method for handling medical dataset with imbalanced classes problem. Adv Math: Sci J 9:6623–6629

Cheng Z, Liu Z, Yang G (2022) Diagnosis of arrhythmia based on multi-scale feature fusion and imbalanced data. In: 2022 7th international conference on machine learning technologies (ICMLT), pp 92–98

Çinaroğlu S (2017) Ensemble learning methods to deal with imbalanced disease and left-skewed cost data. Am J Bioinformat Res 7(1):1–8

Dai D, Hua S (2016) Random under-sampling ensemble methods for highly imbalanced rare disease classification. In: Proceedings of the international conference on data science (ICDATA), p 54

Desuky AS, Omar AH, Mostafa NM (2021) Boosting with crossover for improving imbalanced medical datasets classification. Bull Electr Eng Informat 10(5):2733–2741

Dhanusha C, Kumar AS, Villanueva L (2022) Enhanced contrast pattern based classifier for handling class imbalance in heterogeneous multidomain datasets of Alzheimer disease detection. In: Applications of artificial intelligence and machine learning: select proceedings of ICAAAIML 2021. Springer, pp 801–814

Drosou K, Georgiou S, Koukouvinos C, Stylianou S (2014) Support vector machines classification on class imbalanced data: a case study with real medical data. J Data Sci 12(4):727–753

El-Baz A (2015) Hybrid intelligent system-based rough set and ensemble classifier for breast cancer diagnosis. Neural Comput Appl 26:437–446

Fahmi A, Muqtadiroh FA, Purwitasari D, Sumpeno S, Purnomo MH (2022) A multi-class classification of dengue infection cases with feature selection in imbalanced clinical diagnosis data. Int J Intell Eng Syst 15(3):2022

Farquad MAH, Bose I (2012) Preprocessing unbalanced data using support vector machine. Decis Support Syst 53(1):226–233

Feng Y, Li J (2021) A novel αdistance borderline-adasyn-smote algorithm for imbalanced data and its application in Alzheimer's disease classification based on dense convolutional network. In: Journal of physics: conference series, vol 2031. IOP Publishing, p 012046

Fernando C, Weerasinghe P, Walgampaya C (2022) Heart disease risk iden- tification using machine learning techniques for a highly imbalanced dataset: a comparative study. KDU J Multi Stud 4(2):43–55. https://doi.org/10.4038/kjms.v4i2.50

Fotouhi S, Asadi S, Kattan MW (2019) A comprehensive data level analysis for cancer diagnosis on imbalanced data. J Biomed Inform 90:103089

Fujiwara K, Huang Y, Hori K, Nishioji K, Kobayashi M, Kamaguchi M, Kano M (2020) Over-and undersampling approach for extremely imbalanced and small minority data problem in health record analysis. Front Public Health 8:178

Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F (2011) A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst, Man, Cybern Part C (Appl Rev) 42(4):463–484

Gan D, Shen J, An B, Xu M, Liu N (2020) Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical diagnosis. Comput Ind Eng 140:106266

Gao T, Hao Y, Zhang H, Hu L, Li H, Li H, Hu L, Han B (2018) Predicting pathological response to neoadjuvant chemotherapy in breast cancer patients based on imbalanced clinical data. Pers Ubiquit Comput 22:1039–1047

Ghorbani M, Kazi A, Baghshah MS, Rabiee HR, Navab N (2022) Ra-gcn: graph convolutional network for disease prediction problems with imbalanced data. Med Image Anal 75:102272

Guo H, Liu H, Wu CA, Liu W, She W (2018) Ensemble of rotation trees for imbalanced medical datasets. J Healthc Eng 2018:8902981. https://doi.org/10.1155/2018/8902981

Gupta S, Gupta MK (2022) A comprehensive data-level investigation of cancer diagnosis on imbalanced data. Comput Intell 38(1):156–186

Gupta R, Bhargava R, Jayabalan M (2021) Diagnosis of breast cancer on imbalanced dataset using various sampling techniques and machine learning models. In: 2021 14th international conference on developments in esystems engineering (DeSE). IEEE, pp 162–167

Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G (2017) Learning from class-imbalanced data: Review of methods and applications. Expert Syst Appl 73:220–239

Hallaji E, Razavi-Far R, Palade V, Saif M (2021) Adversarial learning on incomplete and imbalanced medical data for robust survival prediction of liver transplant patients. IEEE Access 9:73641–73650

Han W, Huang Z, Li S, Jia Y (2019) Distribution-sensitive unbalanced data oversampling method for medical diagnosis. J Med Syst 43:1–10

Hassan MM, Amiri N (2019) Classification of imbalanced data of diabetes disease using machine learning algorithms. Age (Years) 21(81):24–33

He F, Yang H, Miao Y, Louis R (2016) A cost sensitive and class-imbalance classification method based on neural network for disease diagnosis. In: 2016 8th international conference on information technology in medicine and education (ITME). IEEE, pp 7–10

Huda S, Yearwood J, Jelinek HF, Hassan MM, Fortino G, Buckland M (2016) A hybrid feature selection with ensemble classification for imbalanced healthcare data: a case study for brain tumor diagnosis. IEEE Access 4:9145–9154

Huo Z, Qian X, Huang S, Wang Z, Mortazavi BJ (2022) Density-aware personalized training for risk prediction in imbalanced medical data. In: Machine learning for healthcare conference. PMLR, pp 101–122

Ibrahim MH (2022) A SALP swarm-based under-sampling approach for medical imbalanced data classification. Avrupa Bilim ve Teknoloji Dergisi 34:396–402

Iori M, Di Castelnuovo C, Verzellesi L, Meglioli G, Lippolis DG, Nitrosi A, Monelli F, Besutti G, Trojani V, Bertolini M et al (2022) Mortality prediction of covid-19 patients using radiomic and neural network features extracted from a wide chest x-ray sample size: A robust approach for different medical imbalanced scenarios. Appl Sci 12(8):3903

Izonin I, Tkachenko R, Greguš M (2022) I-pnn: an improved probabilistic neural network for binary classification of imbalanced medical data. In: Database and expert systems applications: 33rd international conference, DEXA 2022, Vienna, Austria, August 22–24, 2022, Proceedings, Part II. Springer, pp 147–157

Jain A, Ratnoo S, Kumar D (2017) Addressing class imbalance problem in medical diagnosis: a genetic algorithm approach. In: 2017 international conference on information, communication, instrumentation and control (ICICIC). IEEE, pp 1–8

Jain A, Ratnoo S, Kumar D (2023) A novel multi-objective genetic algorithm approach to address class imbalance for disease diagnosis. Int J Info Technol 15:1151–1166. https://doi.org/10.1007/s41870-020-00471-3

Kamaladevi M, Venkatraman V (2021) Tversky similarity based undersampling with Gaussian kernelized decision stump adaboost algorithm for imbalanced medical data classification. Int J Comp Commun Control 16(6):4291. https://doi.org/10.15837/ijccc.2021.6.4291

Kinal M, Woźniak M (2020) Data preprocessing for des-knn and its application to imbalanced medical data classification. In: Intelligent information and database systems: 12th Asian conference, ACIIDS 2020, Phuket, Thailand, March 23–26, 2020, Proceedings, Part I 12. Springer, pp 589–599

Kitchenham B (2004) Procedures for performing systematic reviews. Keele, UK, Keele Univer 33(2004):1–26

Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. Progr Artif Intell 5(4):221–232

Krishnan U, Sangar P (2021) A rebalancing framework for classification of imbalanced medical appointment no-show data. J Data Inf Sci 6(1):178–192

Ksiaa W, Rejab FB, Nouira K (2021) Tuning hyperparameters on unbalanced medical data using support vector machine and online and active svm. In: Intelligent systems design and applications: 20th international conference on intelligent systems design and applications (ISDA 2020) held December 12–15, 2020. Springer, pp 1134–1144

Kumar P, Bhatnagar R, Gaur K, Bhatnagar A (2021) Classification of imbalanced data: review of methods and applications. In: IOP conference series: materials science and engineering, vol 1099. IOP Publishing, p 012077

Kumar V, Medda G, Recupero DR, Riboni D, Helaoui R, Fenu G (2023) How do you feel? Information retrieval in psychotherapy and fair ranking assessment. In: International workshop on algorithmic bias in search and recommendation. Springer, pp 119–133

Kumar P, Thakur RS (2019) Diagnosis of liver disorder using fuzzy adaptive and neighbor weighted k-nn method for lft imbalanced data. In: 2019 international conference on smart structures and systems (ICSSS). IEEE, pp 1–5

Lamari M, Azizi N, Hammami NE, Boukhamla A, Cheriguene S, Dendani N, Benzebouchi NE (2021) Smote–enn-based data sampling and improved dynamic ensemble selection for imbalanced medical data classification. In: Advances on smart and soft computing: proceedings of ICACIn 2020. Springer, pp 37–49

Lan ZC, Huang GY, Li YP, Rho S, Vimal S, Chen BW (2023) Conquering insufficient/imbalanced data learning for the internet of medical things. Neural Comput Appl 35:22949–22958. https://doi.org/10.1007/s00521-022-06897-z

Lee J, Wu Y, Kim H (2015) Unbalanced data classification using support vector machines with active learning on scleroderma lung disease patterns. J Appl Stat 42(3):676–689

Li Y, Hsu WW, Initiative ADN (2022) A classification for complex imbalanced data in disease screening and early diagnosis. Stat Med 41(19):3679–3695

Lijun L, Tingting J, Meiya H (2018) Feature identification from imbalanced data sets for diagnosis of cardiac arrhythmia. In: 2018 11th international symposium on computational intelligence and design (ISCID), vol 2. IEEE, pp 52–55

Liu N, Koh ZX, Chua ECP, Tan LML, Lin Z, Mirza B, Ong MEH (2014) Risk scoring for prediction of acute cardiac complications from imbalanced clinical data. IEEE J Biomed Health Inform 18(6):1894–1902

Liu T, Fan W, Wu C (2019) A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artif Intell Med 101:101723

Liu N, Li X, Qi E, Xu M, Li L, Gao B (2020) A novel ensemble learning paradigm for medical diagnosis with imbalanced data. IEEE Access 8:171263–171280

Li H, Wang X, Li Y, Qin C, Liu C (2018) Comparison between medical knowledge based and computer automated feature selection for detection of coronary artery disease using imbalanced data. In: BIBE 2018; international conference on biological information and biomedical engineering. VDE, pp 1–4

Li J, Xin B, Yang Z, Xu J, Song S, Wang X (2021) Harmonization centered ensemble for small and highly imbalanced medical data classification. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI). IEEE, pp 1742–1745

López V, Fernández A, García S, Palade V, Herrera F (2013) An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf Sci 250:113–141

Luo H, Liao J, Yan X, Liu L (2021) Oversampling by a constraint-based causal network in medical imbalanced data classification. In: 2021 IEEE international conference on multimedia and expo (ICME). IEEE, pp 1–6

Lv J, Chen X, Liu X, Du D, Lv W, Lu L, Wu H (2022) Imbalanced data correction based pet/ct radiomics model for predicting lymph node metastasis in clinical stage t1 lung adenocarcinoma. Front Oncol 12:61

Lyra S, Leonhardt S, Antink CH (2019) Early prediction of sepsis using random forest classification for imbalanced clinical data. In: 2019 computing in cardiology (CinC). IEEE, pp 1–4

Mathew G, Obradovic Z (2013) Distributed privacy-preserving decision support system for highly imbalanced clinical data. ACM Trans Manag Inf Syst (TMIS) 4(3):1–15

Meher PK, Rao AR, Wahi SD, Thelma B (2014) An approach using random forest methodology for disease risk prediction using imbalanced case-control data in gwas. Curr Med Res Pract 4(6):289–294

Mienye ID, Sun Y (2021) Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. Informat Med Unlocked 25:100690

Mohd F, Abdul Jalil M, Noora NMM, Ismail S, Yahya WFF, Mohamad M (2019) Improving accuracy of imbalanced clinical data classification using synthetic minority over-sampling technique. In: Advances in data science, cyber security and IT applications: 1st international conference on computing, ICC 2019, Riyadh, Saudi Arabia, December 10–12, 2019, Proceedings, Part I. Springer, pp 99–110

Mustafa N, Li JP, Memon RA, Omer MZ (2017) A classification model for imbalanced medical data based on pca and farther distance based synthetic minority oversampling technique. Int J Adv Comput Sci Appl 8(1):61–67

Naghavi N, Miller A, Wade E (2019) Towards real-time prediction of freezing of gait in patients with Parkinson's disease: addressing the class imbalance problem. Sensors 19(18):3898

Nalluri MR, Kannan K, Gao XZ, Roy DS (2020) Multiobjective hybrid monarch butterfly optimization for imbalanced disease classification problem. Int J Mach Learn Cybern 11:1423–1451

Napierala K, Stefanowski J (2016) Types of minority class examples and their influence on learning classifiers from imbalanced data. J Intell Inf Syst 46:563–597

Napierala K, Stefanowski J (2012) Identification of different types of minority class examples in imbalanced data. In: Hybrid artificial intelligent systems: 7th international conference, HAIS 2012, Salamanca, Spain, March 28–30th, 2012. Proceedings, Part II, vol 7. Springer, pp 139–150

Naseriparsa M, Al-Shammari A, Sheng M, Zhang Y, Zhou R (2020) Rsmote: improving classification performance over imbalanced medical datasets. Health Inf Sci Syst 8:1–13

Neocleous AC, Nicolaides KH, Schizas CN (2016) Intelligent noninvasive diagnosis of aneuploidy: raw values and highly imbalanced dataset. IEEE J Biomed Health Inform 21(5):1271–1279

Nguyen HT, Tran TB, Bui QM, Luong HH, Le TP, Tran NC (2020) Enhancing disease prediction on imbalanced metagenomic dataset by cost-sensitive. Int J Adv Comput Sci Appl 11(7):651–3657. https://doi.org/10.14569/IJACSA.2020.0110778

Orooji A, Kermani F (2021) Machine learning based methods for handling imbalanced data in hepatitis diagnosis. Front Health Informat 10(1):57

Parvin H, Minaei-Bidgoli B, Alinejad-Rokny H (2013) A new imbalanced learning and dictions tree method for breast cancer diagnosis. J Bionanosci 7(6):673–678

Patel H, Singh Rajput D, Thippa Reddy G, Iwendi C, Kashif Bashir A, Jo O (2020) A review on classification of imbalanced data for wireless sensor networks. Int J Distrib Sens Netw 16(4):1550147720916404

Phankokkruad M (2020) Cost-sensitive extreme gradient boosting for imbalanced classification of breast cancer diagnosis. In: 2020 10th IEEE international conference on control system, computing and engineering (ICCSCE). IEEE, pp 46–51

Polat K (2018) Similarity-based attribute weighting methods via clustering algorithms in the classification of imbalanced medical datasets. Neural Comput Appl 30:987–1013

Porwik P, Orczyk T, Lewandowski M, Cholewa M (2016) Feature projection k-nn classifier model for imbalanced and incomplete medical data. Biocybern Biomed Eng 36(4):644–656

Potharaju SP, Sreedevi M (2016) Ensembled rule based classification algorithms for predicting imbalanced kidney disease data. J Eng Sci Technol Rev 9(5):201–207

Rahman MM, Davis DN (2013) Addressing the class imbalance problem in medical datasets. Int J Mach Learn Comput 3(2):224

Rath A, Mishra D, Panda G, Satapathy SC (2021) Heart disease detection using deep learning methods from imbalanced ecg samples. Biomed Signal Process Control 68:102820

Rath A, Mishra D, Panda G (2022) Imbalanced ecg signal-based heart disease classification using ensemble machine learning technique. Front Big Data 5:1021518. https://doi.org/10.3389/fdata.2022.1021518

Razzaghi T, Safro I, Ewing J, Sadrfaridpour E, Scott JD (2019) Predictive models for bariatric surgery risks with imbalanced medical datasets. Ann Oper Res 280:1–18

Richter AN, Khoshgoftaar TM (2018) Building and interpreting risk models from imbalanced clinical data. In: 2018 IEEE 30th international conference on tools with artificial intelligence (ICTAI). IEEE, pp 143–150

Rodriguez-Almeida AJ, Fabelo H, Ortega S, Deniz A, Balea-Fernandez FJ, Quevedo E, Soguero-Ruiz C, Wägner AM, Callico GM (2023) Synthetic patient data generation and evaluation in disease prediction using small and imbalanced datasets. IEEE J Biomedi Health Info 27(6):2670–2680. https://doi.org/10.1109/JBHI.2022.3196697

Rong P, Luo T, Li J, Li K (2020) Multi-label disease diagnosis based on unbalanced ecg data. In: 2020 IEEE 9th data driven control and learning systems conference (DDCLS). IEEE, pp 253–259

Roy S, Roy U, Sinha D, Pal RK (2023) Imbalanced ensemble learning in determining Parkinson's disease using keystroke dynamics. Expert Syst Appl 217:119522. https://doi.org/10.1016/j.eswa.2023.119522

Sadrawi M, Sun WZ, Ma MHM, Yeh YT, Abbod MF, Shieh JS (2018) Ensemble genetic fuzzy neuro model applied for the emergency medical service via unbalanced data evaluation. Symmetry 10(3):71

Sajana T, Narasingarao M (2018) Classification of imbalanced malaria disease using Naïve Bayesian algorithm. Int J Eng Technol 7(2.7):786–790

Sajana T, Narasingarao M (2018) An ensemble framework for classification of malaria disease. ARPN J Eng Appl Sci 13(9):3299–3307

Salman I, Vomlel J (2017) A machine learning method for incomplete and imbalanced medical data. In: Proceedings of the 20th Czech-Japan seminar on data analysis and decision making under uncertainty, pp 188–195

Shakhgeldyan K, Geltser B, Rublev V, Shirobokov B, Geltser D, Kriger A (2020) Feature selection strategy for intrahospital mortality prediction after coronary artery bypass graft surgery on an unbalanced sample. In: Proceedings of the 4th international conference on computer science and application engineering, pp 1–7

Shaw SS, Ahmed S, Malakar S, Sarkar R (2021) An ensemble approach for handling class imbalanced disease datasets. In: Proceedings of international conference on machine intelligence and data science applications: MIDAS 2020. Springer, pp 345–355

Shilaskar S, Ghatol A (2019) Diagnosis system for imbalanced multi-minority medical dataset. Soft Comput 23(13):4789–4799

Shilaskar S, Ghatol A, Chatur P (2017) Medical decision support system for extremely imbalanced datasets. Inf Sci 384:205–219

Shi X, Qu T, Van Pottelbergh G, Van Den Akker M, De Moor B (2022) A resampling method to improve the prognostic model of end-stage kidney disease: a better strategy for imbalanced data. Front Med 9:730748. https://doi.org/10.3389/fmed.2022.730748

Silveira ACD, Sobrinho Á, Silva LDD, Costa EDB, Pinheiro ME, Perkusich A (2022) Exploring early prediction of chronic kidney disease using machine learning algorithms for small and imbalanced datasets. Appl Sci 12(7):3673

Špečkauskien ̇eV (2015) Feature selection on imbalanced data set for the decision support of Parkinson's disease. In Biomedical Engineering-2015: Proceedings of 19th International conference:[Kaunas, Lithuania, 26-2 November 2015]/Kaunas University of Technology. Biomedical Engineering Institute. Lithuanian Society of Biomedical Engineering. Kaunas: Technologija, 2015, pp. 10–14

Špečkauskien ̇eV (2011) Development and analysis of informational clinical decision support method. Phd thesis, Technologija, Kaunas

Spelmen VS, Porkodi R (2018) A review on handling imbalanced data. In: 2018 international conference on current trends towards converging technologies (ICCTCT). IEEE, pp 1–11

Sribhashyam S, Koganti S, Vineela MV, Kalyani G (2022) Medical diagnosis for incomplete and imbalanced data. In: Intelligent Data Engineering and Analytics: Proceedings of the 9th international conference on frontiers in intelligent computing: theory and applications (FICTA 2021). Springer, pp 491–499

Sridevi T, Murugan A (2014) A novel feature selection method for effective breast cancer diagnosis and prognosis. Int J Comput Appl 88(11):28–33

Srinivas K, Rao GR, Govardhan A (2014) Adapting rough-fuzzy classifier to solve class imbalance problem in heart disease prediction using fcm. Int J Med Eng Informat 6(4):297–318

Sug H (2016) More balanced decision tree generation for imbalanced data sets including the Parkinson's disease data. Int J Biol Biomed Eng 10:115–123

Sun Y, Wong AK, Kamel MS (2009) Classification of imbalanced data: a review. Int J Pattern Recognit Artif Intell 23(04):687–719

Sun H, Wang A, Feng Y, Liu C (2021) An optimized random forest classification method for processing imbalanced data sets of Alzheimer's disease. In: 2021 33rd Chinese control and decision conference (CCDC). IEEE, pp 1670–1673

Suresh T, Brijet Z, Subha T (2023) Imbalanced medical disease dataset classification using enhanced generative adversarial network. Comput Methods Biomech Biomed Eng 26(14):1702–1718. https://doi.org/10.1080/10255842.2022.2134729

Tang X, Cai L, Meng Y, Gu C, Yang J, Yang J (2021) A novel hybrid feature selection and ensemble learning framework for unbalanced cancer data diagnosis with transcriptome and functional proteomic. IEEE Access 9:51659–51668

Tavares TR, Oliveira AL, Cabral GG, Mattos SS, Grigorio R (2013) Preprocessing unbalanced data using weighted support vector machines for prediction of heart disease in children. In: The 2013 international joint conference on neural networks (IJCNN). IEEE, pp 1–8

Venkatanagendra K, Ussenaiah M (2019) Xgb classification technique to resolve imbalanced heart disease data. Int J Res Electron Comput Eng 7(1):406–410

Vinothini A, Baghavathi Priya S (2020) Design of chronic kidney disease prediction model on imbalanced data using machine learning techniques. Indian J Comput Sci Eng 11(6):708–718

Vuttipittayamongkol P, Elyan E (2020a) Improved overlap-based undersampling for imbalanced dataset classification with application to epilepsy and Parkinson's disease. Int J Neural Syst 30(08):2050043. https://doi.org/10.1142/S0129065720500434

Vuttipittayamongkol P, Elyan E (2020b) Overlap-based undersampling method for classification of imbalanced medical datasets. In: Artificial intelligence applications and innovations: 16th IFIP WG 12.5 international conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II, vol 16. Springer, pp 358–369

Wan X, Liu J, Cheung WK, Tong T (2014) Learning to improve medical decision making from imbalanced data without a priori cost. BMC Med Informat Decis Mak 14:1–9

Wang L, Zhao Z, Luo Y, Yu H, Wu S, Ren X, Zheng C, Huang X (2020) Classifying 2-year recurrence in patients with dlbcl using clinical variables with imbalanced data and machine learning methods. Comput Methods Programs Biomed 196:105567

Wang Y, Wei Y, Yang H, Li J, Zhou Y, Wu Q (2020) Utilizing imbalanced electronic health records to predict acute kidney injury by ensemble learning and time series model. BMC Med Informat Decis Mak 20(1):1–13

Wang X, Ren H, Ren J, Song W, Qiao Y, Ren Z, Zhao Y, Linghu L, Cui Y, Zhao Z et al (2023) Machine learning-enabled risk prediction of chronic obstructive pulmonary disease with unbalanced data. Comput Methods Progr Biomed 230: https://doi.org/10.1016/j.cmpb.2023.107340

Wang J, Yao Y, Zhou H, Leng M, Chen X (2013) A new over-sampling technique based on svm for imbalanced diseases data. In: Proceedings 2013 international conference on mechatronic sciences, electric engineering and computer (MEC). IEEE, pp 1224–1228

Wang Q, Zhou Y, Zhang W, Tang Z, Chen X (2020) Adaptive sampling using self-paced learning for imbalanced cancer data pre-diagnosis. Expert Syst Appl 152:113334. https://doi.org/10.1016/j.eswa.2020.113334

Wei X, Jiang F, Wei F, Zhang J, Liao W, Cheng S (2017) An ensemble model for diabetes diagnosis in large-scale and imbalanced dataset. In: Proceedings of the computing frontiers conference, pp 71–78

Werner A, Bach M, Pluskiewicz W (2016) The study of preprocessing methods' utility in analysis of multidimensional and highly imbalanced medical data. In: Proceedings of 11th international conference IIIS2016

Wilk S, Stefanowski J, Wojciechowski S, Farion KJ, Michalowski W (2016) Application of preprocessing methods to imbalanced clinical data: An experimental study. In: Information technologies in medicine: 5th international conference, ITIB 2016 Kamień Śląski, Poland, June 20–22, 2016 proceedings, vol 1. Springer, pp 503–515

Wosiak A, Karbowiak S (2017) Preprocessing compensation techniques for improved classification of imbalanced medical datasets. In: 2017 Federated conference on computer science and information systems (FedCSIS). IEEE, pp 203–211

Woźniak M, Wieczorek M, Siłka J (2023) Bilstm deep neural network model for imbalanced medical data of iot systems. Futur Gener Comput Syst 141:489–499

Wu JC, Shen J, Xu M, Liu FS (2020) An evolutionary self-organizing cost-sensitive radial basis function neural network to deal with imbalanced data in medical diagnosis. Int J Comput Intell Syst 13(1):1608–1618

Xiao Y, Wu J, Lin Z (2021) Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data. Comput Biol Med 135: https://doi.org/10.1016/j.compbiomed.2021.104540

Xu Z, Shen D, Nie T, Kou Y (2020) A hybrid sampling algorithm combining m-smote and enn based on random forest for medical imbalanced data. J Biomed Informat 107:103465

Xu Z, Shen D, Nie T, Kou Y, Yin N, Han X (2021) A cluster-based oversampling algorithm combining smote and k-means for imbalanced medical data. Inf Sci 572:574–589

Yildirim P (2017) Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction. In: 2017 IEEE 41st annual computer software and applications conference (COMPSAC), vol 2. IEEE, pp 193–198

Yuan X, Chen S, Sun C, Yuwen L (2021) A novel class imbalance-oriented polynomial neural network algorithm for disease diagnosis. In: 2021 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, pp 2360–2367

Zeng M, Zou B, Wei F, Liu X, Wang L (2016) Effective prediction of three common diseases by combining smote with tomek links technique for imbalanced medical data. In: 2016 IEEE international conference of online analysis and computing science (ICOACS). IEEE, pp 225–228

Zhang J, Chen L (2019) Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis. Computer Assist Surg 24(sup2):62–72

Zhang H, Zhang H, Pirbhulal S, Wu W, Albuquerque VHCD (2020) Active balancing mechanism for imbalanced medical data in deep learning-based classification models. ACM Trans Multimedia Comput, Commun, Appl (TOMM) 16(1s):1–15

Zhang J, Chen L (2019a) Breast cancer diagnosis from perspective of class imbalance. Iran J Med Phys 16(3). https://doi.org/10.22038/ijmp.2018.31600.1373

Zhang F, Petersen M, Johnson L, Hall J, O'bryant SE (2022) Hyperparameter tuning with high performance computing machine learning for imbalanced Alzheimer's disease data. Appl Sci 12(13):6670

Zhao YX, Yuan H, Wu Y (2021) Prediction of adverse drug reaction using machine learning and deep learning based on an imbalanced electronic medical records dataset. In: Proceedings of the 5th international conference on medical and health informatics, pp 17–21

Zhao H, Wang R, Lei Y, Liao WH, Cao H, Cao J (2022) Severity level diagnosis of Parkinson's disease by ensemble k-nearest neighbor under imbalanced data. Expert Syst Appl 189:116113

Zhou PY, Wong AK (2021) Explanation and prediction of clinical data with imbalanced class distribution based on pattern discovery and disentanglement. BMC Med Informat Decis Mak 21(1):1–15

Zhu M, Xia J, Jin X, Yan M, Cai G, Yan J, Ning G (2018) Class weights random forest algorithm for processing class imbalanced medical data. IEEE Access 6:4641–4652

Zięba M (2014) Service-oriented medical system for supporting decisions with missing and imbalanced data. IEEE J Biomed Health Informat 18(5):1533–1540