Artificial Intelligence Review (2024) 57:80
https://doi.org/10.1007/s10462-023-10652-8

Cost-sensitive learning for imbalanced medical data: a review

Imane Araf[1] · **Ali Idri**[1,2] · **Ikram Chairi**[3]

Published online: 1 March 2024

Abstract

Integrating Machine Learning (ML) in medicine has unlocked many opportunities to harness complex medical data, enhancing patient outcomes and advancing the field. However, the inherent imbalanced distribution of medical data poses a significant challenge, resulting in biased ML models that perform poorly on minority classes. Mitigating the impact of class imbalance has prompted researchers to explore various strategies, wherein Cost-Sensitive Learning (CSL) arises as a promising approach to improve the accuracy and reliability of ML models. This paper presents the first review of CSL for imbalanced medical data. A comprehensive exploration of the existing literature encompassed papers published from January 2010 to December 2022 and sourced from five major digital libraries. A total of 173 papers were selected, analysed, and classified based on key criteria, including publication years, channels and sources, research types, empirical types, medical sub-fields, medical tasks, CSL approaches, strengths and weaknesses of CSL, frequently used datasets and data types, evaluation metrics, and development tools. The results indicate a noteworthy publication rise, particularly since 2020, and a strong preference for CSL direct approaches. Data type analysis unveiled diverse modalities, with medical images prevailing. The underutilisation of cost-related metrics and the prevalence of Python as the primary programming tool are highlighted. The strengths and weaknesses analysis covered three aspects: CSL strategy, CSL approaches, and relevant works. This study serves as a valuable resource for researchers seeking to explore the current state of research, identify strengths and gaps in the existing literature and advance CSL's application for imbalanced medical data.

Keywords Machine learning · Data imbalance · Cost-sensitive learning · Medical data

✉ Ali Idri
ali.idri@um5.ac.ma

Imane Araf
imane.araf@um6p.ma

Ikram Chairi
ikram.chairi@um6p.ma

[1] Faculty of Medical Sciences, Mohammed VI Polytechnic University (UM6P), Ben Guerir, Morocco

[2] ENSIAS, Mohammed V University, Rabat, Morocco

[3] College of Computing, Mohammed VI Polytechnic University (UM6P), Ben Guerir, Morocco

1 Introduction

Machine Learning (ML) techniques have gained significant attention in medicine (Rajkomar et al. 2019), offering promising opportunities to enhance diagnostic accuracy, improve treatment outcomes, and optimise healthcare delivery. However, the unique characteristics of medical data present substantial challenges that must be addressed to harness the full potential of ML in healthcare.

One such challenge is the issue of class imbalance, which arises when the distribution of classes in a dataset is highly skewed. In medical datasets, imbalanced distributions are frequently observed, where the prevalence of certain medical conditions is significantly lower than others. For instance, in the context of postoperative risk evaluation, and considering the short planning period of one year, the number of patients surviving the expected duration is often significantly higher than the number of deceases recorded (Zieba et al. 2014). Traditional ML algorithms often struggle to handle these imbalances, leading to biased models that perform poorly on the minority class, which is more often than not the class of interest.

Addressing class imbalance is essential for building reliable and effective ML models in the medical domain. Many strategies have been proposed in the literature, including resampling (Khushi et al. 2021), ensemble learning (Galar et al. 2012), and Cost-Sensitive Learning (CSL) (Elkan 2001). Resampling techniques aim to rebalance the class distribution in the dataset by oversampling the minority class or undersampling the majority class. Rebalancing holds the potential to improve the performance of the models significantly. However, it is important to consider the limitations of resampling. Oversampling may lead to overfitting, where the model becomes overly specialised to the minority class, while undersampling may result in the loss of valuable information from the majority class (Hu et al. 2021). Ensemble learning, on the other hand, combines multiple models to improve overall performance. Ensemble-based methods can be tailored to tackle class imbalance by incorporating resampling or CSL (López et al. 2013; Fernández et al. 2018). However, adopting ensemble learning can introduce computational complexity (Galar et al. 2012), requiring additional resources and time for training and inference.

In contrast, CSL offers an alternative strategy that preserves the data distribution while ensuring computational efficiency. CSL introduces distinct misclassification costs for each class. The underlying assumption is that higher misclassification costs are assigned to samples from the minority class, and the objective is to minimise the high-cost errors (López et al. 2013). This strategy is advantageous in numerous real-world scenarios, particularly medical applications, where certain misclassifications can have more severe consequences (Sterner et al. 2021). For example, mislabelling a cancer patient as healthy is more detrimental than the opposite scenario, as it can result in delayed treatment and further complications.

Despite the growing interest in CSL for medical research, the existing literature remains fragmented and lacks comprehensive studies that provide a systematic overview of the field. Previous reviews (Sterner et al. 2021; Freitas et al. 2009) suffer from limitations such as a lack of systematic approach, outdatedness, or limited scope, hindering the development of a clear and up-to-date understanding of CSL's application to imbalanced data in medicine. This paper systematically reviews the use of CSL for imbalanced medical data, marking the first study of its kind to the best of our knowledge. Our study entails a thorough review of peer-reviewed papers sourced from reputable databases. Through a systematic search process, meticulous analysis of pertinent literature, and extraction of key findings, our objective is to provide valuable insights and practical guidance to researchers in this particular domain and suggest potential future research directions. Extensive exploration was undertaken to comprehensively investigate the existing literature, covering the period from January 2010 to December 2022. Five major digital libraries were meticulously searched to collect relevant materials, including PubMed, ScienceDirect, IEEE Xplore, SpringerLink, and Google Scholar. The 173 selected papers were subsequently analysed to answer nine Research Questions (RQs): (i) publication years, channels, and sources; (ii) research types; (iii) empirical types; (iv) medical disciplines; (v) medical tasks; (vi) CSL approaches; (vii) strengths and weaknesses of CSL, (viii) frequently used datasets, data types, and evaluation metrics; and (iv) development tools.

The remainder of this paper is organised as follows. Section 2 delves into the background of class imbalance and introduces the fundamental concepts of CSL. Section 3 presents the research methodology employed in this study. Section 4 provides a detailed analysis of the derived statistical trends. In Sect. 5, a comprehensive overview of CSL approaches is presented. Section 6 offers an in-depth analysis of the strengths and weaknesses of CSL, along with a comparative assessment of selected works. Section 7 focuses on datasets and data types. Subsequently, Sect. 8 describes the performance metrics used to evaluate CSL techniques. Section 9 covers the development tools employed for CSL techniques' implementation. Section 10 discusses the limitations of this study. The implications of the results and practical guidance for researchers are presented in Sect. 11, emphasising the key takeaways and actionable recommendations for future investigations. Finally, Sect. 12 concludes the paper and outlines future research directions.

2 Background

This section introduces two key concepts: class imbalance and CSL. We aim to define and establish a clear understanding of these concepts, offering background information on the challenges posed by class imbalance in medical data classification. Additionally, we will delve into the fundamental principles and considerations of CSL, setting the stage for further exploration in subsequent sections.

2.1 The class imbalance problem

Class imbalance is a prevalent phenomenon observed in many real-world datasets, where the distribution of instances across different classes is significantly skewed. While the technical definition of class imbalance encompasses any dataset with unequal class distributions, the term typically refers to datasets that exhibit substantial and sometimes extreme imbalances (He and Garcia 2009). This imbalance has garnered considerable attention from researchers and practitioners due to its widespread occurrence in various classification problems, such as anomaly detection (Zhou et al. 2021), face recognition (Huang et al. 2020), medical diagnosis (Mazurowski et al. 2008), and more. In such scenarios, the minority class, often referred to as the positive class, represents the concept of interest and is characterised by its low frequency. On the other hand, the majority class, also known as the negative class, constitutes the class with higher representation. The scarcity of instances belonging to the minority class can stem from their inherent exceptional or rare nature, or it may result from the high cost of acquiring data for these particular examples (López et al. 2013). Consequently, accurately identifying and classifying instances from the minority class becomes crucial, as it often carries significant implications in practical applications.

While class imbalance is commonly associated with binary classification tasks, it is crucial to note that imbalanced distributions can extend to multi-class (Wang and Yao 2012) and multi-label (Lankireddy et al. 2022) settings. Within such contexts, the presence of multiple minority classes introduces additional complexities, amplifying the challenge at hand (López et al. 2013; Lankireddy et al. 2022). Moreover, it is essential to acknowledge that class imbalance is not limited to classification tasks alone but can also manifest in segmentation tasks, specifically in medical imaging (Rezaei et al. 2019). In such scenarios, a significant disparity emerges between the number of pixels or voxels representing regions of interest and those corresponding to the background class or other classes (Nasalwai et al. 2021).

The degree of class imbalance in a dataset can be quantified using the Imbalance Ratio (IR), which is calculated as the ratio of the number of examples in the majority class N_{maj} to the number of examples in the minority class N_{min}:

$$IR = \frac{N_{\text{maj}}}{N_{\text{min}}} \tag{1}$$

The IR metric quantifies the severity of class imbalance, providing insights into dataset composition and aiding in devising effective strategies to address imbalanced datasets. An annotation such as 1:100 can be used to represent an IR of 100, indicating that the majority class is approximately 100 times more prevalent than the minority class.

2.2 Strategies to mitigate class imbalance

Mitigating class imbalance is crucial due to several key reasons. Primarily, biased learning poses a significant challenge. Standard learning algorithms are designed to work optimally on balanced datasets, where the numbers of instances in each class are roughly equal. When applied to imbalanced datasets, these algorithms tend to be biased towards the majority class, leading to suboptimal classification models and frequent misclassification of minority class instances (Fernández et al. 2018). Consequently, the minority class instances are frequently misclassified, reducing overall performance. Moreover, the significance of the minority class in various real-world scenarios cannot be overstated. Misclassifying samples from the minority class can have severe ramifications, such as missed opportunities, erroneous diagnoses, or potential risks. Furthermore, the natural occurrence of imbalanced datasets due to the inherent characteristics of the problem domain adds another layer of complexity. For instance, rare medical conditions often have limited data availability compared to more prevalent cases. To address these challenges, it is imperative to develop effective strategies that specifically target class imbalance. By doing so, classification performance can be significantly improved, ensuring accurate identification of minority class instances and enabling the extraction of valuable insights from the available imbalanced datasets.

Many strategies have been proposed in the literature to address the class imbalance challenge. These strategies can be classified into four groups (Fernández et al. 2018; Haixiang et al. 2017):

- Data-level strategies, also known as external strategies, focus on modifying the dataset to rebalance the class distribution. Techniques such as oversampling, undersampling, and hybrid methods are employed. Oversampling methods generate synthetic examples or replicate existing instances of the minority class to augment its representation. Conversely, undersampling methods reduce the number of examples from the majority class to achieve a more balanced dataset. Hybrid methods combine oversampling and undersampling techniques to achieve the desired class distribution. These modifications are typically performed as a preprocessing step to ensure improved model performance (Fernández et al. 2018).

- Algorithm-level strategies, also called internal strategies, involve adapting the learning algorithms to assign greater importance to the minority class. These strategies require a deeper understanding of the model and the application domain to identify why the model fails when under imbalanced class distributions (Fernández et al. 2018).
- Cost-sensitive strategy considers the varying costs associated with misclassifications across different classes. This strategy lies between data-level and algorithm-level strategies. They can operate at the data level by assigning costs to individual instances or at the algorithm level by incorporating cost considerations into the learning process (López et al. 2013; Fernández et al. 2018).
- Ensemble-based strategies combine multiple base learners to create a more accurate and robust classification model. These strategies can be adapted to handle imbalanced datasets in two ways. Firstly, the ensemble learning algorithm can be modified at the data level, enabling preprocessing steps to be performed on the data before the learning stage of each classifier (López et al. 2013; Fernández et al. 2018). Alternatively, a cost-sensitive framework can be incorporated to build cost-sensitive ensembles. Rather than altering the base classifier to accept costs during the learning process, cost-sensitive ensembles are designed to guide the cost minimisation procedure through the ensemble learning algorithm (López et al. 2013; Fernández et al. 2018). Galar et al. (2012) present a comprehensive taxonomy of ensemble methods for learning with imbalanced classes in their review. The authors predominantly categorise these ensemble strategies into four distinct families. The first family encompasses cost-sensitive boosting methods, while the remaining three families incorporate data preprocessing techniques and are further classified based on the ensemble learning algorithm employed, namely boosting, bagging, and hybrid ensembles.

For a closer examination of these strategies, Table 1 provides a breakdown of their strengths and weaknesses, which have been carefully curated from prior reviews and discussions on addressing class imbalance. These strategies enhance model performance and effectively tackle class imbalance in various contexts. Data-level strategies offer a promising starting point due to their ease of implementation, straightforwardness, flexibility, and versatility, as they remain independent of the underlying algorithm. However, they are not without their trade-offs. Oversampling techniques risk overfitting and often demand extended training times, while undersampling methods can potentially discard informative samples from the majority class. In contrast, algorithm-level strategies introduce targeted solutions to class imbalance without altering the underlying data distribution. They exhibit an advantage in that they are less likely to affect training time, yet they demand an in-depth understanding of the algorithm in use and are inherently algorithm-specific, potentially compromising their flexibility and ease of implementation. CSL stands out for its computational efficiency and data distribution preservation while addressing class imbalance. Nonetheless, it presents challenges in setting appropriate misclassification costs, often requiring careful optimisation and facing potential overfitting to the minority class during cost tuning. Note that a detailed exposition of the strengths and weaknesses associated with CSL is provided in Subsection 6.1. Ensemble-based

Table 1 Strengths and weaknesses of balancing strategies

Strategy	Strengths	Studies	Weaknesses	Studies
Data-level strategies (oversampling)	Easy implementation	Haixiang et al. (2017); Kaur et al. (2019); Rekha et al. (2019)	Potential overfitting	López et al. (2013); He and Garcia (2009); Kaur et al. (2019); Rekha et al. (2019); Elrahman and Abraham (2013); Feng et al. (2020); Patel et al. (2020); Leevy et al. (2018); Johnson and Khoshgoftaar (2019)
	Versatile (Independent of the algorithm)	López et al. (2013); Kaur et al. (2019); Elrahman and Abraham (2013); Tarekegn et al. (2021)	Longer training time	Elrahman and Abraham (2013); Feng et al. (2020); Patel et al. (2020); Leevy et al. (2018); Johnson and Khoshgoftaar (2019)
	Straightforward	Kaur et al. (2019)		
Data-level strategies (undersampling)	Flexible	Kaur et al. (2019); Rekha et al. (2019)	Loss of information	López et al. (2013); He and Garcia (2009); Kaur et al. (2019); Rekha et al. (2019); Elrahman and Abraham (2013); Feng et al. (2020); Patel et al. (2020); Leevy et al. (2018)
Algorithm-level strategies	Do not cause any shifts in the data distribution	Fernández et al. (2018); Johnson and Khoshgoftaar (2019)	Require a deep understanding of the algorithm	Fernández et al. (2018); Kaur et al. (2019); Tarekegn et al. (2021)
	More directed alleviation of the imbalance problem	Fernández et al. (2018)	Reduced flexibility	Fernández et al. (2018); Tarekegn et al. (2021)
	Less likely to impact training time	Johnson and Khoshgoftaar (2019)	More difficult to design and implement than data-level strategies	Fernández et al. (2018)

Table 1 (continued)

Strategy	Strengths	Studies	Weaknesses	Studies
CSL	Computationally efficient	Haixiang et al. (2017); Kaur et al. (2019); Tarekegn et al. (2021); Johnson and Khoshgoftaar (2019)	Misclassification costs are unknown	López et al. (2013); He and Garcia (2009); Haixiang et al. (2017); Kaur et al. (2019); Rekha et al. (2019); Elrahman and Abraham (2013); Tarekegn et al. (2021); Feng et al. (2020); Patel et al. (2020); Leevy et al. (2018); Johnson and Khoshgoftaar (2019)
	Preserves the data distribution	Kaur et al. (2019)	Risk of overfitting when searching for optimal costs	Kaur et al. (2019); Elrahman and Abraham (2013)
Ensemble-based strategies	Multiple classifiers provide better prediction than a single classifier	Rekha et al. (2019); Elrahman and Abraham (2013); Leevy et al. (2018)	The output model can be difficult to interpret	López et al. (2013)
	More resilience to noise (decreased variance)	Elrahman and Abraham (2013)	Computational complexity	López et al. (2013); Rekha et al. (2019); Elrahman and Abraham (2013); Tarekegn et al. (2021)
	Improved generalizability	Elrahman and Abraham (2013; Tarekegn et al. (2021)		

strategies, leveraging multiple classifiers, offer superior predictive performance and improved resilience to noise, thereby enhancing generalizability. However, they come with the trade-off of reduced model interpretability and increased computational complexity. Considering these diverse strategies, researchers and practitioners can select the most suitable solution based on the specific problem's requirements and constraints, as each strategy offers a unique blend of advantages and considerations to guide their decision-making process.

2.3 CSL

This section introduces the core concepts of CSL and presents an illustrative example of cost-sensitive logistic regression applied to a cervical cancer diagnosis dataset.

2.3.1 Overview

CSL encompasses a group of algorithms designed to account for varying misclassification costs associated with False Positives (FP) and False Negatives (FN). The CSL concept becomes particularly relevant in the medical field when examining the consequences of such misclassifications. For example, in the context of prognosis, specifically in predicting the risk of recurrence in cancer patients, misclassifying patients with a high risk of recurrence as low risk is more costly and dangerous than the opposite scenario. Such misclassification can lead to inadequate surveillance and delayed interventions, resulting in increased chances of disease progression, complications, and higher healthcare costs. On the other hand, misclassifying patients with a low risk of recurrence as high risk may result in unnecessary tests, which can still incur additional expenses but with less immediate harm to the patient. While misclassification costs are the primary focus of this paper, it is worth mentioning that other types of costs (Turney 2002), such as attribute costs (Uguroglu et al. 2012), can also be incorporated into the learning process.

CSL has gained significant attention in addressing uneven class distributions. Nevertheless, it is essential to note that CSL is not limited to imbalanced scenarios but also finds application in balanced datasets where misclassifications can have severe outcomes (Fernández et al. 2018). While resampling is more commonly employed in imbalanced data settings, CSL offers distinct advantages regarding computational efficiency (Haixiang et al. 2017). Furthermore, many empirical studies have showcased the superiority of cost-sensitive techniques over resampling techniques in some application domains (He and Garcia 2009).

The efficacy of CSL heavily depends on the supplied cost matrix, which quantifies the costs $C(i, j)$ associated with misclassifying samples from one class j as another class i. Table 2 provides an illustrative example of a cost matrix for a binary classification scenario. These cost values can be determined by domain experts or estimated using training data (Fernández et al. 2018; Ling and Sheng 2008). Notably, cost attribution assumes a higher penalty for

Table 2 Cost matrix for a binary classification scenario

	Actual negative	Actual positive
Predicted negative	C(0,0)	C_p=C(0,1)
Predicted positive	C_n=C(1,0)	C(1,1)

misclassifying a positive instance compared to a negative one, while correct classifications incur no costs ($C(0,0) = C(1,1) = 0$).

By leveraging the cost matrix, the classification of a given example is guided by the minimum expected cost principle (Elkan 2001; López et al. 2013; Fernández et al. 2018; Ling and Sheng 2008). Accordingly, the example is classified into the class with the lowest expected cost. The expected cost (conditional risk) $R(i \mid x)$ of classifying an example x into class i can be formulated as:

$$R(i \mid x) = \sum_{j} P(j \mid x) \cdot C(i,j) \tag{2}$$

Here, $P(j \mid x)$ represents the probability estimate of classifying an example x as belonging to class j. In the context of binary classification, a cost-sensitive classifier will classify an example x into the positive class if and only if the following condition holds true:

$$P(0 \mid x) \cdot C(1,0) + P(1 \mid x) \cdot C(1,1) \leq P(0 \mid x) \cdot C(0,0) + P(1 \mid x) \cdot C(0,1)$$

This condition can be equivalently restated as:

$$P(0 \mid x) \cdot (C(1,0) - C(0,0)) \leq P(1 \mid x) \cdot (C(0,1) - C(1,1))$$

Considering the assumption that $C(0,0) = C(1,1) = 0$, the classifier will classify an example x as belonging to the positive class if and only if:

$$P(0 \mid x) \cdot C(1,0) \leq P(1 \mid x) \cdot C(0,1)$$

As $P(0 \mid x) = 1 - P(1 \mid x)$, a threshold p^* can be derived for classifying an instance x into the positive class if $P(1 \mid x) \geq p^*$, where:

$$p^* = \frac{C(1,0)}{C(1,0)+C(0,1)} \tag{3}$$

2.3.2 Illustrative example

In this subsection, we present an illustrative example designed to provide a clear and insightful demonstration of how CSL can be practically applied in medical data analysis. Our focus centres on diagnosing cervical cancer, employing the Cervical Cancer Risk Factors dataset (Fernandes et al. 2017). Here, the correct identification of cancer cases is prioritised, recognising its greater importance and critical nature compared to identifying non-cancer patients.

The Cervical Cancer Risk Factors dataset comprises 858 instances, with 803 categorised as healthy individuals and 55 diagnosed with cervical cancer, resulting in a significant class imbalance with an IR of approximately 1:15. This dataset includes 32 distinct features related to medical history, habits, and demographic information, all associated with the risk factors leading to biopsy examinations for cervical cancer.

Before applying CSL, we performed essential data preprocessing steps, including handling missing values and feature scaling. These steps ensure the dataset is appropriately prepared for model training, although detailed elaboration falls beyond the primary focus of our discussion. We also split the dataset into training and test sets, allocating 80% for training and 20% for testing. The details of this dataset splitting are summarised in Table 3 below.

Table 3 Cervical cancer risk factors dataset splitting details

	Total instances	Cancer instances	Non-cancer instances
Training set	686	44	642
Test set	172	11	161

Table 4 Cost matrix for the illustrative example on cervical cancer diagnosis

	Actual non-cancer	Actual cancer
Predicted non-cancer	0	15
Predicted cancer	1	0

Our construction of the cost matrix follows a common practice in CSL. We set the misclassification cost for the majority class (non-cancer cases) to 1 and for the minority class (cancer cases) to the IR. The cost matrix is presented in Table 4, and further details on selecting misclassification costs are discussed in Subsection 6.1.

For modelling, we opted for logistic regression for its simplicity and wide use in medical research. The following pseudo-code (Algorithm 1) outlines the steps for training the logistic regression model with the cost-sensitive strategy to account for class imbalance:

Algorithm 1 Logistic regression training

```
Initialise logistic regression model
for each training instance $i$ do
    Extract features $X_i$ and true label $y_i$
    if $y_i$ corresponds to a diagnosis of cervical cancer then
        Assign a higher misclassification cost (15)
    else
        Assign a lower misclassification cost (1)
    end if
end for
Train the model with all training instances $(X, y)$ and their assigned costs
```

In our evaluation, we conducted a comparative analysis of cost-insensitive and cost-sensitive logistic regression to assess the impact of CSL on diagnostic accuracy. We employed standard evaluation metrics: accuracy, precision, sensitivity (recall), F1 score and the Area Under the Receiver Operating Characteristic Curve (AUC). These metrics allow us to comprehensively evaluate the model's ability to distinguish cervical cancer from non-cancer cases. The results of this evaluation are presented in Table 5. For further insights about CSL evaluation, we refer readers to the detailed discussion in Sect. 8.

Applying CSL to our model resulted in substantial improvements in key performance metrics. Sensitivity increased substantially from 54.5% to 81.8%, allowing us to correctly identify 27.3% more cancer cases. This enhancement is pivotal in medical diagnostics, as it minimises the risk of missing cases. Moreover, precision improved from 50% to 56.2%,

Table 5 Comparative analysis of cost-insensitive and cost-sensitive logistic regression for cervical cancer diagnosis

Metric	Cost-insensitive model	Cost-sensitive model
Accuracy	93.6%	94.8%
Precision	50%	56.2%
Sensitivity	54.5%	81.8%
F1 score	52.2%	66.6%
AUC	89.2%	95.4%

indicating increased accuracy in positive predictions and fewer false alarms. The F1 score rose from 52.2% to 66.6%, achieving a better balance between missed cases and false alarms. The AUC also increased from 89.2% to 95.4%, demonstrating the model's heightened ability to distinguish cancer from non-cancer cases. While there was a slight improvement in accuracy, rising from 93.6% to 94.8%, this demonstrates a modest yet valuable boost in overall classification correctness.

In essence, CSL elevated our model's cancer diagnosis capabilities and refined its precision, making it a valuable tool for cervical cancer diagnosis. It is important to note that this analysis represents an illustrative simple example, and CSL's benefits can extend further in more complex medical scenarios.

3 Methodology

The present study follows the guidelines proposed by Petersen et al. (2015). The process covers: (i) clearly defining the RQs, (ii) developing a comprehensive search strategy to identify relevant papers, (iii) screening the identified papers based on inclusion and exclusion criteria, (iv) designing a classification scheme, and (v) data extraction and analysis.

3.1 Research questions

This study aims to provide an overview and a structured understanding of the existing literature on using CSL for imbalanced medical data. To this end, nine RQs were identified and are presented along with their rationales in Table 6.

3.2 Search strategy

The search is conducted in five digital libraries: PubMed, ScienceDirect, IEEE Xplore, SpringerLink, and Google Scholar from January 2010 until December 2022. These libraries were chosen based on their extensive coverage of peer-reviewed publications in the fields of medicine and health sciences, as well as computer science and engineering.

The search string was formulated based on the principal terms from the RQs and the PICO (Population, Intervention, Comparison, and Outcomes) framework (Kitchenham and Charters 2007). Note that the third and fourth letters of PICO were not included in the search string formulation since neither empirical comparison nor measurable outcomes were considered in this study. Additionally, the search string was expanded to

Table 6 Research questions

ID	RQ	Rationale
RQ1	In which years, publication channels, and sources were the selected papers published?	To explore the historical publication trends and identify the different channels and sources in which the selected papers were published
RQ2	What types of research were published?	To determine the research types presented in the selected studies
RQ3	Which empirical methods are used to evaluate cost-sensitive models in medicine?	To examine the types of empirical validation performed to evaluate cost-sensitive models in medicine
RQ4	In which disciplines of medicine was CSL mainly employed?	To determine the medical disciplines in which CSL was applied
RQ5	Which medical tasks are addressed in the selected papers?	To identify the medical tasks for which CSL was used
RQ6	Which CSL approaches were most frequently used in medicine?	To identify the most frequent CSL approaches in the medical literature
RQ7	What are the strengths and weaknesses of cost-sensitive methods in medicine?	To point out the strengths and limitations of cost-sensitive techniques in medicine
RQ8	What are the frequently used medical datasets, data types, and metrics to assess the performance of cost-sensitive models?	To determine the most employed datasets, data types, and metrics to assess the performance of cost-sensitive models in medicine
RQ9	Which development tools are used for cost-sensitive techniques’ implementation?	To identify the development tools employed to implement cost-sensitive techniques

include alternative spellings and synonyms of the derived terms to ensure a comprehensive search.

The main search terms were initially linked with their substitutes using the Boolean operator "OR" and were joined using "AND" afterwards. Table 7 exhibits the complete search string, where the "scope" column demonstrates the main terms and the "search terms" column displays the related keywords.

3.3 Study selection

The Inclusion Criteria (IC) and Exclusion Criteria (EC) utilised to identify the relevant papers are presented in Table 8. The systematic selection process involved a multi-tiered approach, as outlined below:

1. *Initial Screening*: The evaluation of titles, abstracts, and keywords of papers obtained from the selected databases was initiated. This initial review facilitated the exclusion of papers that were clearly irrelevant to the study, thereby streamlining the candidate pool.
2. *Extended Review*: In cases where uncertainty about a paper's relevance remained after the initial screening, a more thorough examination was undertaken. This involved reviewing the paper's introduction, discussion, and conclusion sections to help determine whether it should be included in the study.
3. *Full-Text Review*: Full-text reading was selectively performed when the information obtained during the extended review was insufficient to decide on a paper's relevance to the study.

One author conducted the initial examination of the papers, and the remaining authors evaluated the final selection. Any disagreements during this process were resolved through constructive discussions in meetings, ultimately leading to a consensus on the final set of included studies.

Table 7 Search string

Scope	Search terms
Medicine	Health* *OR* Medic* *OR* Disease *OR* Clinic*
AND Artificial Intelligence	"Machine Learning" *OR* "Deep Learning" *OR* Intelligen* *OR* Classif* *OR* Predict* *OR* Diagnos* *OR* Prognos*
AND Technique	Technique *OR* Method *OR* Tool *OR* Model *OR* Algorithm *OR* Approach *OR* Framework
AND CSL	"Cost sensitive" *OR* Cost-sensitive *OR* "weighted cost function" *OR* "weighted loss function" *OR* "class weighting" *OR* re-weighting
AND Imbalance	Imbalance* *OR* unbalance* *OR* "skewed class distribution" *OR* under-represented *OR* "majority class" *OR* "minority class"

Table 8 Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
IC1: Studies developing new or using existing cost-sensitive techniques in medicine	EC1: Papers published earlier than January 2010 or later than December 2022
IC2: Papers focusing mainly on cost-sensitive models in medicine, whether or not comparing them to other balancing techniques	EC2: Papers using several datasets from multiple areas with a mere presence of medical ones
IC3: Papers presenting fair comparisons of several balancing techniques in medicine, including cost-sensitive methods	EC3: Papers using cost-sensitive techniques in public health, biology, pharmacology, or genomics
IC4: Papers presenting comparisons between CSL methods in medicine without proposing any newly developed techniques	EC4: Papers available as abstracts, posters, book chapters (excluded due to potential duplication with previously published conference or journal papers), or presentations
IC5: Papers providing an overview of studies investigating cost-sensitive methods in medicine	EC5: Non-peer-reviewed papers
IC6: Papers combining cost-sensitive methods with other balancing techniques in medicine	EC6: Duplicate publications of the same study
	EC7: Studies published in languages other than English
	EC8: Short papers
	EC9: Papers for which the full texts are not available

3.4 Quality assessment

While Quality Assessment (QA) remains optional in the guidelines followed, it is still recommended (Petersen et al. 2015) as it can help minimise the risk of bias, guide the interpretation of findings, and determine the strength of inferences (Kitchenham and Charters 2007), thereby improving the overall validity of the study.

Each paper was evaluated by two authors based on the checklist in Table 9 to ensure that the selected studies are of sufficient quality and provide reliable and valid evidence to address the RQs. The checklist comprised six QA criteria, but only QA5 and QA6 were deemed relevant for reviews and theoretical papers. The decision to exclude the first four criteria was based on their applicability to empirical studies only, and their use for reviews and theoretical papers may have compromised their eligibility for inclusion in this study.

3.5 Data extraction strategy and synthesis

In this phase, a data extraction form was developed to retrieve relevant information from the selected papers, addressing the RQs in Table 6. The structure of the form is detailed in Table 10. One author conducted the task meticulously, while the other two authors rigorously reviewed the extracted data to ensure its accuracy and objectivity.

Data synthesis aims to summarise and synthesise the extracted data pertaining to each RQ. The vote-counting method is utilised to aid in result interpretation, followed by a narrative synthesis to comprehensively report and discuss the outcomes for each

Table 9 QA checklist

ID	Questions	Possible answers and scoring
QA1	Does the study give clear empirical results?	Yes (+1), No (+0)
QA2	Does the study give a justified empirical design?	Yes (+1), No (+0), Partially (+0.5)
QA3	Does the study evaluate the performance of the developed solution?	Yes (+1), No (+0), Partially (+0.5)
QA4	Is the proposed solution in the study compared to other solutions?	Yes (+1), No (+0)
QA5	Does the study explicitly present the proposed method's benefits and limitations?	Yes (+1), No (+0), Partially (+0.5)
QA6	Is the study published in a recognised source?	For conferences and workshops: Core2021: A/A* (+1.5), B (+1), C (+0.5), No Rank (+0) For journals: JCR2021: Q1 (+2), Q2 (+1.5), Q3 (+1), No Rank (+0)

question. Additionally, visual representations in the form of charts and tables are incorporated to enhance the clarity and presentation of the findings.

3.6 Study selection results

Figure 1 displays the number of articles at each stage of the selection process. Initially, 49325 candidate papers were identified, from which 49124 studies were discarded according to the IC and EC. 28 studies that did not fulfil the QA criteria were later excluded. Eventually, 173 papers were retained to answer the RQs. The list of selected papers and their extracted data can be obtained through an email request to the authors.

4 Statistical trends

This section presents a detailed analysis of the statistical trends observed among the selected studies.

4.1 Publication trends

Figure 2 shows the number of selected studies per publication channel from January 2010 to December 2022. Three main channels were identified: journals, conferences, and workshops. Out of the 173 selected studies, the majority, precisely 69.9% (121 papers), were published in journals, 27.2% (47 papers) were published in conference proceedings, and only 2.9% (five papers) were published in workshops. Table 11 outlines the publication sources that have published more than two papers. The findings indicate that Computer Methods and Programs in Biomedicine was the most commonly targeted journal venue, while the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) emerged as the most frequently occurring source for conference papers. Chronologically speaking, conference papers were the dominant publication type

Table 10 Data extraction form

Study identifier
Title
Publication year
Authors
Abstract
Digital library
RQ1: *In which years, publication channels, and sources were the selected papers published?*
Publication years, channels (journal, conference, or workshop), and sources were extracted to address this question.
RQ2: *What types of research were published?*
The research types were categorised as follows: evaluation research, validation research, solution proposal, review, and others (philosophical papers, opinion papers, and experience papers) (Petersen et al. 2015).
RQ3: *Which empirical methods are used to evaluate cost-sensitive models in medicine?*
The empirical methods can be classified as historical-based evaluation, case study, or survey (Petersen et al. 2015).
RQ4: *In which disciplines of medicine was CSL mainly employed?*
Each paper was examined to determine its specific medical focus, encompassing disciplines such as oncology, cardiology, ophthalmology, and others, as detailed exhaustively in (Careers in medicine 2023).
RQ5: *Which medical tasks are addressed in the selected papers?*
The medical tasks can be classified into screening, diagnosis, prognosis, treatment, monitoring, and management (Esfandiari et al. 2014).
RQ6: *Which CSL approaches were most frequently used in medicine?*
The developed cost-sensitive methods in the selected studies were identified. These methods can be classified as either direct or meta-learning approaches. The latter could further be classified as preprocessing or postprocessing methods (Fernández et al. 2018).
RQ7: *What are the strengths and weaknesses of cost-sensitive methods in medicine?*
The strengths and weaknesses of CSL, CSL approaches, and some selected works were outlined.
RQ8: *What are the frequently used medical datasets, data types, and metrics to assess the performance of cost-sensitive models?*
The frequently used medical datasets, data types (numeric, categorical, time series, images, or text), and evaluation metrics were retrieved.
RQ9: *Which development tools are used for cost-sensitive techniques' implementation?*
The reported development tools (programming language, package, or software) were identified.

in 2012 and 2013. However, the trend shifted in 2014 as the journal publication frequency surpassed that of conference papers in subsequent years. A key observation is that the gap between the two types of publications became increasingly pronounced from 2020 onwards. The analysis further revealed a growing trend of publications, particularly since 2020, when the count peaked significantly. Notably, no study was published in 2010, and only one workshop paper was published in 2011.

The dearth of published papers in 2010–2011 and the dominance of conference papers until 2013 suggest that CSL research in the medical field was in its early stages. However, as the field progressed, researchers began to prioritise top-tier journals due to their strict review processes and higher publication standards, resulting in more rigorous research. This shift towards journal publications started in 2014 when the number of journal articles surpassed conference papers and continued to widen in subsequent years. This trend

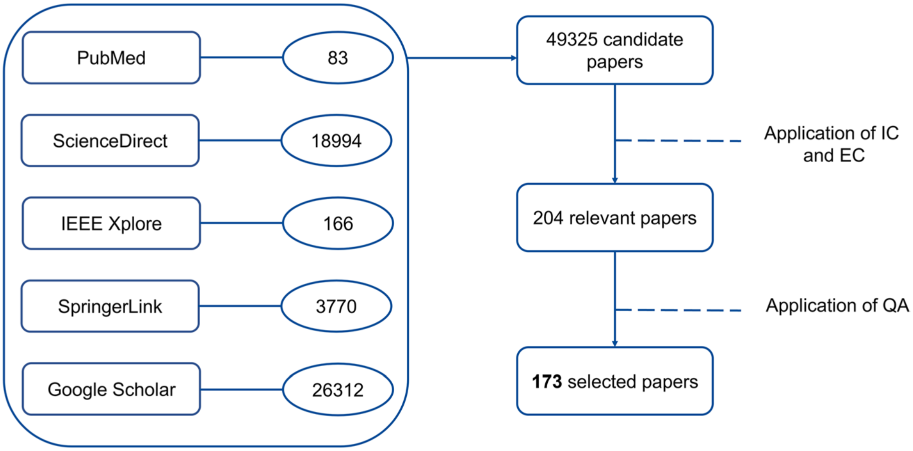

Fig. 1 Selection process

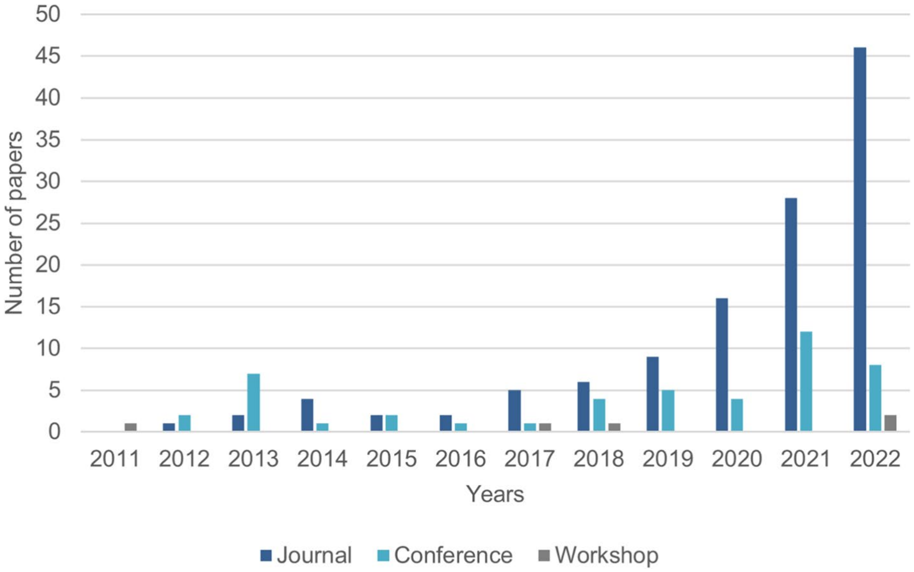

Fig. 2 Distribution of the selected papers per publication year and channel

indicates a maturing field and researchers increasingly meeting the demanding standards of high-quality journals.

The growing interest and abundance of publications on CSL can be attributed to several key factors. Firstly, the development of high-throughput technologies has resulted in massive amounts of medical data (Johnson et al. 2018), including clinical data, electronic health records, and data from wearable devices. These advancements in data collection have created an urgent need for novel methods to analyse and leverage this data for improved medical outcomes. Secondly, the inherent imbalanced nature of this collected

Table 11 Publication sources

Journal source	#Papers	Percentage
Computer Methods and Programs in Biomedicine	9	5.2%
Computers in Biology and Medicine	8	4.6%
BMC Medical Informatics and Decision Making	5	2.9%
Neurocomputing	5	2.9%
Multimedia Tools and Applications	5	2.9%
Medical Image Analysis	4	2.3%
Biomedical Signal Processing and Control	4	2.3%
Artificial Intelligence in Medicine	3	1.7%
Applied Soft Computing	3	1.7%
Other	75	43.4%
Conference source	**#Papers**	**Percentage**
International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)	5	2.9%
Other	42	24.3%
Workshop source	**#Papers**	**Percentage**
International Workshop on Machine Learning in Medical Imaging (MLMI)	3	1.7%
Other	2	1.2%

data poses a critical challenge that impacts the accuracy and reliability of ML models in medical applications. Thirdly, the significant advances in CSL algorithms (Khan et al. 2018) and their success in other fields (Sahin et al. 2013) have encouraged researchers to apply these techniques in the medical domain, where they are much needed. Additionally, the advances in Deep Learning (DL) techniques have been a significant catalyst for progress in medical data analysis (Esteva et al. 2019). Finally, the increasing availability of public datasets and tools for analysing medical data has facilitated the dissemination and replication of research findings. As a result, the research community has become more aware of the importance of addressing the imbalance problem, leading to a surge in publications on this topic, particularly in recent years.

Besides, the findings revealed diverse publication sources covering various disciplines such as medicine, medical informatics, computer science, and artificial intelligence. This diversity reflects the interdisciplinary nature of the research topic, requiring a multi-faceted approach that draws on expertise from different fields.

4.2 Research types

After scrutinising the selected studies, four distinct research types were identified: Evaluation Research (ER), Validation Research (VR), Solution Proposal (SP), and reviews. No other research types were observed. Of the studied literature, 147 papers (85%) were found to be both SP and ER, introducing new or improved cost-sensitive methods and testing them on medical data. ER was the second most frequent type, comprising 23 studies (13.3%), whereas VR was the focus of only two papers (1.2%) published in the years 2018 and 2021 (Wang et al. 2018a; Aldraimli et al. 2021). Notably, one paper (0.6%) in 2021 stood out as a combined review and ER effort, initially surveying existing methods before conducting performance benchmarking (Rahman et al. 2021b).

The evolution over time of the two most frequent research types, ER and SP, is displayed in Fig. 3. It is apparent from the line chart that the number of SP with ER was consistently higher than that of ER alone. The number of papers proposing new cost-sensitive techniques or enhancing existing ones rose from 2011 to 2013, peaking at six papers per year before falling in 2014 and levelling off at three papers until 2016. After that, SP+ER studies surged significantly, especially after 2020, with the highest number of papers (52) published in 2022. Conversely, the trend for studies evaluating existing solutions followed a different pattern. They first appeared in 2012 with one paper, peaked at three papers in 2013, and then declined to zero papers in 2016, where they remained until 2019, except for one paper published in 2017. In 2020, five ER studies resurfaced, increasing slightly to six studies in 2021 and then falling to four studies in 2022.

The analysis revealed that all the papers proposing new CSL methods also conducted experimental evaluations to demonstrate their effectiveness. This is a noteworthy point, as it indicates that researchers are not simply proposing theoretical solutions but are also committed to demonstrating the practical value of their work.

Furthermore, the dominance of SP+ER papers in the literature suggests that researchers primarily focus on proposing new methods for CSL rather than evaluating existing solutions. While this could be attributed to the complexity of the CSL problem and the unique challenges posed by imbalanced medical data, it also indicates the significant investment and interest in advancing the state-of-the-art in this area. The surge of SP+ER studies after 2020 suggests an increasing awareness of the importance of effective CSL methods in medical applications. This trend is further fueled by advancements in ML, the availability of larger datasets, and increasing computational resources,

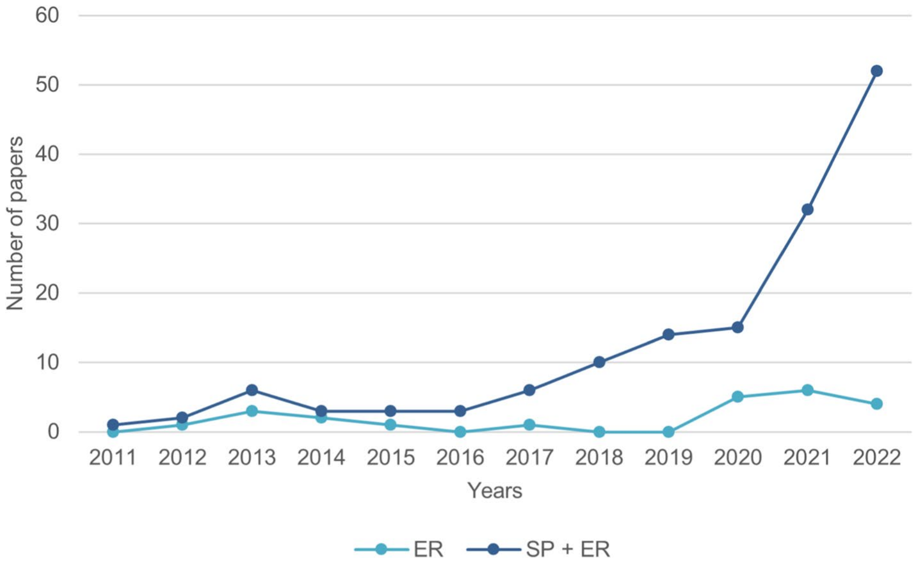

Fig. 3 Evolution of research types per year

enabling researchers to develop more effective CSL techniques and conduct more extensive and complex experiments.

Out of all the selected papers, only two focused on validating cost-sensitive methods, which may be attributed to the challenges associated with conducting validation studies in hospital settings. Such studies require close collaboration with medical professionals and access to sensitive patient data. Nevertheless, the limited number of validation studies suggests a need for further research to demonstrate and validate the practical value of CSL in real-world medical settings. Besides, the fact that there is only one review paper in the literature points to the necessity for more synthesis and critical evaluation of existing CSL methods in medicine.

Overall, the findings suggest a promising outlook for the future of CSL for medical data but also underscore the need for continued validation and rigorous evaluation of the developed techniques.

4.3 Empirical types

The selected studies were evaluated using three empirical methods: Case Study (CS), Historical-Based evaluation (HBE), and survey. Figure 4 illustrates the distribution of research and empirical types. It can be observed from the bubble plot that HBE was the most prevalent empirical type, with 117 papers (65.4%) using publicly available medical datasets to assess their models. Of these papers, the majority (105) proposed and evaluated novel or improved solutions, 11 studies evaluated existing ones, and only one study was dedicated to reviewing and evaluating previously suggested methods. The CS empirical type ranked second, with 60 papers (33.5%) using real-world datasets from hospitals or healthcare units. Among these papers, 46 were classified as SP studies, 12 as ER studies, and two as VR studies. By contrast, survey-based evaluations were relatively uncommon, with only two SP studies (1.1%) employing this method. It is worth noting that six papers used both public and real-life datasets and were hence double-counted in HBE and CS empirical types. Besides, 12 CS papers used real-world data from their previous works.

The prevalence of HBE studies indicates that many scholars rely on existing, publicly available datasets to assess their models. This practice partly owes to the abundance and ease of accessibility of historical data. However, it also stems from the challenges

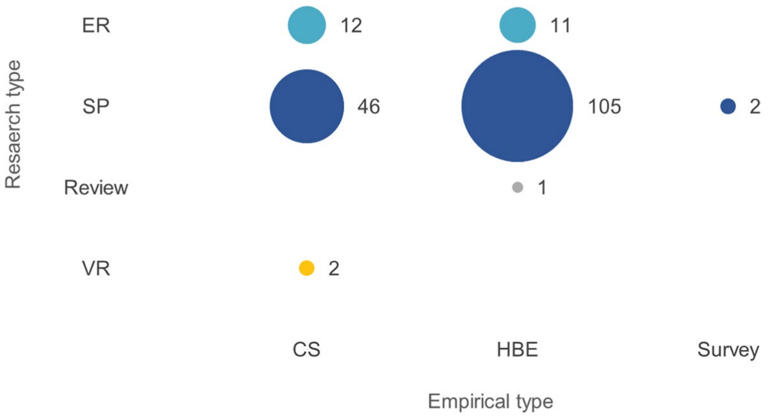

Fig. 4 Distribution of the selected studies per research and empirical types

associated with obtaining real-world medical data, including ethical and legal considerations (Mello et al. 2018). HBE is considered a cost-effective and accessible way to evaluate models on a large scale and benchmark novel solutions against existing literature. Nonetheless, it is essential to recognise that HBE may not accurately reflect the complexities and nuances of real-world medical data. Therefore, researchers may need to supplement their HBE findings with other empirical methods, such as CS.

A less yet significant number of papers used CS with real-world data collected from healthcare units or hospitals. This indicates that researchers are keenly interested in testing their models in practical settings to ensure their applicability. CS provides a more detailed and nuanced understanding of how models perform in specific contexts and can prove valuable in validating or fine-tuning models developed using publicly available datasets. However, it is worth acknowledging that CS is typically more resource-intensive than HBE and requires collaboration with medical professionals and institutions.

Combining public and real-life medical datasets demonstrates an understanding of the strengths and limitations of each. Four papers merged these two types of data and employed the CS type for tasks such as collecting healthy controls (Wang et al. 2018b), gathering additional negative samples (Calderon-Ramirez et al. 2021), or testing solutions after training on historical data (Pranto et al. 2020). While this approach may enhance models' generalizability and practical applicability, it may also introduce complexity by requiring additional preprocessing to ensure data comparability and consistency, as highlighted by Wang et al. (2018b). Therefore, researchers using this approach should be transparent about their methodology. The remaining two studies (Hu et al. 2021; Xu et al. 2020) employed the datasets separately to assess the generalizability of their methods. By testing their models on datasets with diverse characteristics and features, researchers can assess how well their models perform in various settings and contexts, contributing to their research's overall reliability and validity.

The relatively limited number of survey-based evaluations may be attributed to the inherent challenges associated with effectively designing and implementing surveys and the multiple sources of bias (Cunningham et al. 2015) that may arise. These sources of bias include, among others: non-response bias, which occurs when patients who choose not to participate in the survey are systematically different from those who do participate; social desirability bias, which stems from respondents providing answers they perceive to be socially desirable rather than truthful; recall bias, which arises from patients inaccurately recalling past events or experiences, such as the duration and timing of symptoms or treatments; and instrument bias, which can occur if the survey instrument itself is flawed or biased, such as when a question is phrased confusingly, potentially distorting the accuracy of responses.

4.4 Medical disciplines

The 173 selected studies collectively explored 21 distinct medical disciplines. Interestingly, 17 papers addressed more than one discipline, either by investigating a topic at the intersection of two medical sub-fields (e.g., (Sung et al. 2021)) or by testing their methods on a diverse range of disciplines (e.g., (Gan et al. 2020)). Figure 5 showcases the distribution of studies per medical sub-field, focusing solely on sub-fields addressed by at least 2% of the selected papers.

The findings revealed that oncology emerged as the discipline garnering the highest degree of attention, accounting for 31.2% (54 papers) of the selected studies. As per the World Health Organization (WHO), cancer is a leading cause of mortality globally,

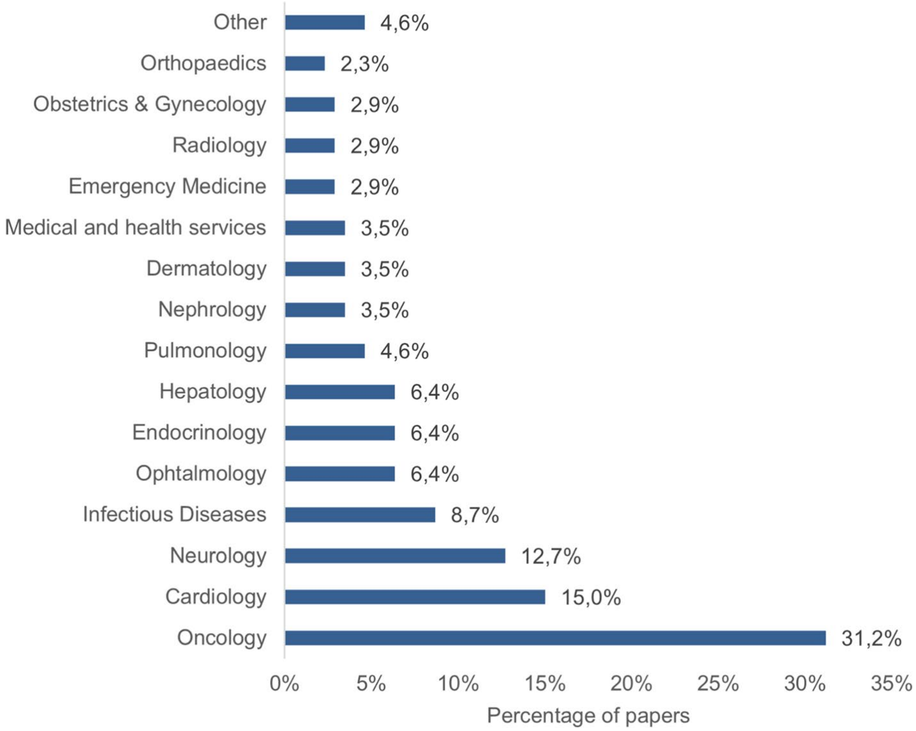

Fig. 5 Distribution of the selected papers per medical discipline

accounting for approximately 10 million deaths in 2020 alone (World Health Organization 2022). The significance of accurate and timely diagnosis and treatment is paramount, and ML techniques hold great promise in this regard. However, cancer is a highly heterogeneous disease that can manifest differently in each patient. Additionally, patients often present with complex medical histories and comorbidities, which can complicate diagnosis and treatment. These factors can contribute to imbalanced medical data, making CSL an attractive approach to address these challenges and improve cancer care.

Cardiology and neurology received significant focus in subsequent order, constituting 15% (26 papers) and 12.7% (22 papers) of the investigated literature, respectively. CSL has demonstrated significant benefits in addressing cardiovascular and neurological diseases, widely recognised as significant health concerns. This finding is in line with the WHO's report (2021), which identifies cardiovascular diseases as the primary cause of mortality globally, responsible for 17.9 million deaths in 2019. Additionally, the WHO acknowledges that neurological disorders such as stroke, Alzheimer's disease, and other dementias are among the leading causes of disability and death worldwide (World Health Organization 2016). Given the high mortality rate associated with these diseases, accurate predictions are imperative. However, data imbalance can lead to biased models that fail to capture important patterns in the data. By adopting CSL, researchers aim to improve prediction accuracy and contribute to preserving human life.

Infectious diseases occupy the fourth position, representing 8.7% (15 papers) of the total studies. Notable attention has been dedicated to researching this sub-field since 2020. This trend is not surprising, considering the urgency and global impact of the COVID-19 pandemic, which first emerged in 2019 and has since garnered substantial research attention. Additionally, imbalanced data is a common issue in COVID-19 studies due to various factors such as differences in testing availability and criteria, variations in reporting standards, differences in demographics, healthcare infrastructure, and compliance with public health measures. Besides, there may be a publication bias towards COVID-19 studies due to the pandemic's global impact, and funding agencies may have prioritised research on this topic. Lastly, data availability may have contributed to the popularity of COVID-19 as a research subject.

Other medical sub-fields, such as ophthalmology, endocrinology, and hepatology, were investigated by 11 papers (6.8%) each, demonstrating the relevance of cost-sensitive methods in these domains. Galdran and colleagues (2020) highlighted the value of cost-sensitive classifiers in addressing two critical challenges in diabetic retinopathy grading. These classifiers can effectively model the complex structure of a heterogeneous label space and are also advantageous in addressing severely class-imbalanced scenarios. Fan et al. (2022) pointed out the inadequacy of conventional models in considering the imbalanced distribution of diabetic datasets and the varying misclassification costs across distinct patient categories. In a previous study by Yang et al. (2021), the predictive accuracy of traditional ML methods and cost-sensitive models were compared for predicting hepatic encephalopathy in cirrhotic patients. The study's results demonstrated the superiority of cost-sensitive models, underscoring their high suitability and potential for future prognosis studies.

Pulmonology was featured in 8 articles (4.6%), and nephrology, dermatology, and medical and health services were each investigated by six studies (3.5%). On the other hand, emergency medicine (2.9%), radiology (2.9%), and obstetrics & gynecology (2.9%) received relatively little attention, as did orthopaedics, which was addressed by only 2.3% of the selected studies (four papers).

Disciplines that received the least amount of attention in the selected studies were classified as "other," which included geriatric psychiatry and neonatology, each addressed by two papers (1.2%), as well as intensive care, radiomics, urology, and podiatry, which were each the focus of only one study (0.6%). This may be explained by factors such as limited data availability and researchers prioritising other research areas deemed more crucial and pertinent to patient care.

4.5 Medical tasks

Upon rigorous analysis, it was observed that all six predefined medical tasks, namely screening, diagnosis, prognosis, management, monitoring, and treatment, were covered in the selected literature. Notably, a small subset of seven papers delved into multiple medical tasks, owing to their utilisation of diverse datasets associated with distinct objectives.

The distribution of studies per medical task is graphically presented in Fig. 6, providing a clear overview of the prevalence of each task within the selected studies. The findings unveiled diagnosis as the most extensively explored medical task, dominating the literature with an overwhelming majority of 66.5% (115 papers). This dominance can be attributed to a multitude of factors. Foremost, diagnosis is the keystone of patient care and treatment decisions, guiding healthcare professionals in determining appropriate therapeutic interventions. Accurate and timely diagnosis allows for identifying the most suitable treatment strategies (Mirbabaie et al. 2021), thereby increasing survival prospects and enhancing patient well-being. Recognising this fundamental role, researchers and practitioners invest significant efforts in developing effective and accurate diagnostic models and algorithms. Moreover, the availability of diverse and well-annotated datasets specifically designed for

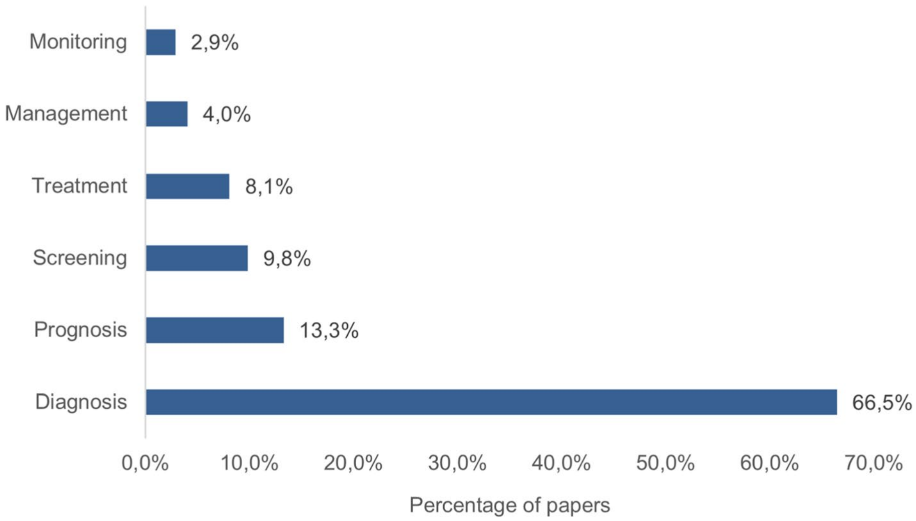

Fig. 6 Distribution of the selected papers per medical task

diagnostic purposes contributes to the preponderance of diagnosis-focused studies. These datasets serve as invaluable resources for training and evaluating diagnostic algorithms, encompassing various medical conditions and their associated diagnostic information. Furthermore, the relatively lower frequency of certain medical conditions within the population results in a scarcity of positive instances. This inherent imbalance within diagnostic datasets further accentuates the significance of exploring and advancing diagnostic methodologies. Additionally, the diagnostic process itself can introduce imbalances in datasets due to the requirement of invasive procedures or expensive tests to confirm certain cases. As a result, researchers are actively exploring CSL techniques tailored to address the challenges posed by imbalanced diagnostic datasets.

Prognosis ranked second in terms of research focus, accounting for 13.3% (23 papers) of the selected studies. The presence of a substantial body of research focused on prognostic prediction underscores its significance in medical research. The study of prognosis holds paramount importance in understanding and predicting the future course of various medical conditions (Moons et al. 2009). It provides healthcare professionals with vital insights into potential outcomes, recovery rates, disease progression, and possible complications. This information enables informed decision-making regarding treatment options, care plans, and patient management strategies. Furthermore, the emphasis on prognosis aligns with the contemporary shift towards precision medicine and patient-centred care (König et al. 2017). By customising interventions based on individual predictions, healthcare providers can improve patient outcomes and enhance the overall quality of care.

Screening was represented by 9.8% (17 papers) of the selected studies, highlighting the importance of early detection in the medical domain. Researchers have recognised the importance of developing effective techniques and models to identify individuals at risk or needing further diagnostic evaluation. Furthermore, the presence of 14 papers (8.1%) dedicated to treatment highlights the efforts invested in optimising therapeutic interventions and evaluating their effectiveness. These papers explore various treatment modalities, including pharmacological interventions (e.g. Wu et al. 2022), surgical procedures (e.g. Dorado-Moreno et al. 2017) and non-pharmacological approaches (e.g. Aldraimli et al. 2022).

On the other hand, management was featured in a limited number of seven papers (4%). The relatively scant attention given to this particular medical task may be attributed to various factors. One potential explanation is that management often involves intricate and multi-faceted strategies, which necessitate a combination of clinical expertise, patient engagement, and healthcare system considerations, which may be challenging to capture solely through data-driven approaches. Additionally, the focus on other medical tasks, namely diagnosis, prognosis, screening, and treatment in the selected studies, reflects the immediate priorities in medical research, where there is substantial stress on improving diagnostic accuracy, predicting clinical results, and optimising treatment interventions. However, despite the scarcity of selected papers on management, it remains a critical aspect of healthcare research. Effective management strategies can significantly impact long-term patient outcomes, quality of care, and resource allocation (Esfandiari et al. 2014). The paucity of publications highlights the need for future investigations and multidisciplinary collaborations to address the complexities of managing medical conditions.

The monitoring task exhibited the lowest representation within selected research, with a modest inclusion of only five papers (2.9%). This disparity in attention can be attributed to the immediate impact that other medical tasks hold, which often overshadows the perception of monitoring as a complementary aspect of care rather than a primary focus. Moreover, the findings may be elucidated by considering data and research resources availability. Monitoring requires longitudinal data collection and continuous observation of patients (Esfandiari et al. 2014), which can be challenging and resource-intensive. Researchers may face constraints when seeking access to large-scale, high-quality monitoring data, resulting in fewer studies in this area. Nonetheless, the limited number of papers on monitoring does not diminish its importance in healthcare. Monitoring assumes a crucial function in evaluating treatment efficacy, identifying early signs of complications, and ensuring patient safety (Khan et al. 2016). Moving forward, future research needs to consider the significance of monitoring in providing comprehensive patient care. Furthermore, researchers should explore innovative approaches to address the challenges encountered in monitoring within medical research.

5 CSL approaches

This section extensively examines CSL approaches, providing detailed explanations for each approach and highlighting their prevalence across the selected research papers.

5.1 Overview

CSL techniques can be broadly classified into two categories: direct approaches and meta-learning approaches (Fernández et al. 2018; Liu et al. 2021; Johnson and Khoshgoftaar 2019; Ling and Sheng 2008; Sheng and Ling 2006). The former category modifies the learning algorithms by incorporating misclassification costs during the model training phase (Fernández et al. 2018; Feng et al. 2020). Conversely, the latter category does not alter the learning algorithms per se (Liu et al. 2021). Instead, meta-learning adjusts the training data (preprocessing) or the model's outputs (postprocessing) to ensure cost sensitivity. Popular preprocessing techniques include instance weighting based on a cost matrix and MetaCost (Fernández et al. 2018), which relabels the training data according to misclassification costs. Postprocessing techniques, meanwhile, often involve adjusting the decision thresholds based on the predefined costs (Fernández et al. 2018; Liu et al. 2021).

5.1.1 Direct approaches

The fundamental concept behind developing a direct cost-sensitive algorithm involves directly incorporating misclassification costs into the underlying learning algorithm. This integration is designed to elevate the significance of the positive class. As a result, the optimisation process transitions from minimising total error to minimising total cost (Johnson and Khoshgoftaar 2019). Numerous research efforts within the literature have explored direct approaches, yielding multiple cost-sensitive adaptations of conventional algorithms. One such algorithm, decision trees (Ling et al. 2004), has seen extensive utilisation in prior studies.

In their work, Ling and colleagues (2004) presented a cost-sensitive modification of decision trees that considers both attribute and misclassification costs with equal importance. This consideration enables the algorithm to handle data imbalance while minimising the feature-related costs. We present a simplified summary of the procedure through a six-step process as follows:

1. Data preprocessing:
    - (a) Discretise numerical attributes if necessary
    - (b) Assign cost values for FP and FN (C_n and C_p)
2. Attribute selection and splitting:
    - (a) Calculate the total cost of splitting based on test cost and misclassification cost
    - (b) Choose the attribute that minimises the total cost as the splitting attribute
    - (c) If attribute costs are non-zero, select attributes that can improve predictive accuracy while minimising the cost
3. Handling unknown attribute values:
    - (a) Treat missing values as a special category
    - (b) Do not build leaves or sub-trees for instances with unknown values
    - (c) Keep examples with unknown values within the node representing the attribute
4. Leaf labelling:
    - (a) At each leaf node, determine whether it should be labelled positive or negative based on cost minimisation
    - (b) Compare the cost of predicting negative (FP) with predicting positive (FN): if $(N_p \times C_p) > (N_n \times C_n)$, label the leaf as positive; otherwise, label it as negative. Here, N_p and N_n represent the number of positive and negative instances in the leaf node, respectively.

5. Tree expansion:

   (a) Continue the process recursively for examples falling into branches of the splitting attribute
   (b) If the cost of splitting further is not beneficial, stop building sub-trees and create a leaf node

6. Overfitting control (optional): While the algorithm does not incorporate tree pruning in its basic form, consider adding post-tree pruning procedures to simplify the tree if needed. Pruning can help control overfitting in scenarios where the tree becomes too complex.

In the scope of direct approaches, other adaptations have emerged beyond decision trees. For instance, a straightforward method can be employed to make K-Nearest Neighbors (KNN) cost-sensitive (Qin et al. 2013; Zhang 2020). In this method, standard KNN is used for training a classifier. When predicting the class for a test sample, k nearest neighbours are selected from the training data. The class probabilities are estimated by considering the ratio of the number of neighbours from each class to the total number of neighbours:

$$P(i|x) = \frac{k_i}{k} \tag{4}$$

where k_i is the number of k nearest neighbours for class i. By adhering to the minimum expected cost principle and employing Eqs. 2 and 4, the optimal class label for each test sample can be computed straightforwardly. Other examples include Support Vector Machines (SVM) (Iranmehr et al. 2019), artificial neural networks (Kukar and Kononenko 1998), Naïve Bayes (Di Nunzio 2014) and random forest (Devi et al. 2019), among several others.

Another aspect of direct approaches involves new cost-sensitive loss functions that enable the minority samples to contribute more to the loss (Johnson and Khoshgoftaar 2019). These specialised loss functions are designed to address class imbalance by assigning higher penalties to the misclassification of minority class instances. Several popular cost-sensitive loss functions have been proposed in the literature. One notable example is the weighted cross-entropy loss (Naceur et al. 2020; Rahman et al. 2021b; Punn and Agarwal 2021), a modification of the standard cross-entropy loss function used in classification problems. The weighted cross-entropy loss can be expressed as:

$$L = -\sum_{i=1}^{K} C_i y_i \log(P(i|x)) \tag{5}$$

where K is the number of classes, C_i is the cost associated with class i, and y_i is the true label for class i.

Additionally, the Focal loss function was initially proposed by Lin et al. (2020) for object detection tasks, where positive foreground samples are significantly outnumbered by negative background samples (Johnson and Khoshgoftaar 2019).

Focal loss provides a dynamic weighting scheme that downplays easily classified instances and emphasises hard-to-classify ones, effectively giving more significance to minority class samples. This is accomplished by multiplying the cross-entropy loss by a scaling factor, $\alpha_i(1 - P(i|x))^\gamma$. The hyperparameter γ controls the extent to which easy examples are de-emphasised, while α_i serves as a class-specific weight to increase the importance of the minority class (Johnson and Khoshgoftaar 2019; Lee et al. 2023; Lu et al. 2021; Li et al. 2022b).

In addition to the weighted cross-entropy loss and Focal loss functions, other cost-sensitive loss functions proposed in the literature include Dice loss (Shirokikh et al. 2020; Taghanaki et al. 2019), asymmetric similarity loss Shirokikh et al. (2020), weighted hinge loss (Wu et al. 2022), and accelerated Tversky loss (Nasalwai et al. 2021).

5.1.2 Instance weighting

In contrast to direct approaches, instance weighting presents a different way of addressing misclassification costs. In this method, greater emphasis is placed on positive instances with higher misclassification costs by assigning them higher weights. Notably, instance weighting operates as a preprocessing solution, diverging from resampling, as it preserves the size of the original training set. To ensure a clear exposition of how cost-sensitivity is achieved through instance weighting, its utilisation within decision trees as an example is expounded upon.

Distinct from the direct incorporation of costs into split creation, Ting (2002) proposed a simple method for assigning instance weights to induce cost-sensitive trees. This method can be seamlessly applied to any existing tree learning. The procedure is as follows (Fernández et al. 2018; Ting 2002):

1. Initially, the cost matrix must be transformed into a cost vector for each class. The conversion formula proposed by Breiman et al. (1984) is employed:

$$C(j) = \sum_{i}^{I} C(i,j) \tag{6}$$

   where $C(i, j)$ is the cost of misclassifying an instance from class j as belonging to class i, and I is the number of classes.
2. Next, the weight of class j is calculated as:

$$w(j) = \frac{C(j)N}{\sum_i C(i)N_i} \tag{7}$$

   In the context of this equation, N represents the total number of instances within the training set, N_i signifies the count of instances belonging to class j, and the summation of all instance weights can be expressed as $\sum_j w(j)N_j = N$.When $C(j) \geq 1$, the weight $w(j)$ assumes its minimum value $0 < \frac{N}{\sum_i C(i)N_i} \leq 1$ when $C(j) = 1$ and reaches its maximum value $w(j) = \frac{C(j)\sum_i N_i}{\sum_i C(i)N_i} \geq 1$ when $C(j) = \max_i C(i)$.
3. The following equation is employed to derive the ratio of the total weight of instances belonging to class j to the total weight in node t:

$$p_w(j|t) = \frac{W_j(t)}{\sum_i W_i(t)} = \frac{w(j)N_j(t)}{\sum_i w(i)N_i(t)} \tag{8}$$

4. Any chosen training procedure for constructing decision trees can be applied without alterations, with the sole adjustment being the substitution of $W_j(t)$ for $N_j(t)$ when calculating the splitting criterion value at each node during the tree growth process, as well as in the error estimation in the pruning process.

5.1.3 MetaCost

MetaCost is known for its versatility in making classifiers cost-sensitive. This method acts as a wrapper, compatible with various classifier types, regardless of their output, be it class labels or probability estimates (Fernández et al. 2018; Domingos 1999). MetaCost, as introduced by Domingos (1999), can be conceptually dissected into three phases: Private ensemble building, relabelling, and classification (Siers and Islam 2020). MetaCost starts by creating multiple bootstrap samples from the initial training set following the bagging ensemble method, and each of these samples is used to train individual classifiers. These classifiers are then aggregated, either through averaging if the classifier used produces class probabilities or through a majority-voting scheme, to determine the probabilities of each example belonging to different classes. The original training examples in the dataset are subsequently relabelled to minimise the conditional risk defined in Equation 2. The resulting relabelled training data are then utilised to train the final classifier.

The pseudo-code (Domingos 1999) for the MetaCost procedure is provided in Algorithm 2.

Algorithm 2 MetaCost algorithm

```
Require: $S$: training set, $L$: classification algorithm, $C$: cost matrix, $m$: number of
  bootstrap samples to generate, $n$: size of the bootstrap sample
  for $i = 1$ to $m$ do
      Let $S_i$ be a sample of $S$ with $n$ examples
      Let $M_i$ be the model produced by applying $L$ to $S_i$
  end for
  for each instance $x$ in $S$ do
      for each class $j$ do
          if $L$ produces class probabilities then
              Obtain $P(j|x, M_i)$
          else
              $P(j|x, M_i) = 1$ for the class predicted by $M_i$ for $x$, and 0 for all others
          end if
          if all bootstrap samples are to be used for each instance then
              $i$ ranges over all $M_i$
          else
              $i$ ranges over all $M_i$ such that $x \notin S_i$
          end if
          $P(j|x) = \frac{1}{\sum_i 1} \quad \sum_i P(j|x, M_i)$
      end for
      Let the class of $x = \arg\min_i \sum_j P(j|x)C(i, j)$
  end for
  Let $M$ be the model produced by applying $L$ to $S$
  Return $M$
```

5.1.4 Thresholding

Thresholding is a postprocessing technique employed in CSL to fine-tune the classification decisions of a trained model based on specific cost considerations. Unlike other approaches, thresholding operates on the output probability estimates produced by a model after training during the test phase (Fernández et al. 2018; Johnson and Khoshgoftaar 2019; Vanderschueren et al. 2022). It serves as a meta-learning approach, allowing the conversion of any cost-insensitive model into a cost-sensitive one (Johnson and Khoshgoftaar 2019; Sheng and Ling 2006), and possesses the advantage of being more accessible and comprehensible to practitioners (Feng et al. 2020).

Thresholding operates by minimising the expected cost based on the specified cost matrix. It employs the threshold p* defined in Eq. 3 to adapt the decision threshold (usually set at 0.5 in conventional classification) when categorising samples in a way that reduces bias towards the majority class (Johnson and Khoshgoftaar 2019; Sheng and Ling 2006). Owing to its versatility, thresholding has found practical utility in a wide range of algorithms in the literature (Liu et al. 2021; Sheng and Ling 2006; Zhou and Liu 2006; Zhang and Shen 2011; Cao et al. 2013a; Zhou et al. 2014; Zhao 2008), consistently demonstrating positive performance results.

In addition to the technique described above, another method for threshold optimisation is empirical thresholding (Zhao et al. 2018; Reychav et al. 2019). This method involves iteratively searching for the optimal threshold that minimises the total cost on a validation set (Vanderschueren et al. 2022), offering an alternative means of optimising cost-sensitive classification.

5.2 The distribution of CSL approaches in the selected studies

This study seeks to categorise the selected papers according to the cost-sensitive approaches they have employed, with the goal of obtaining a thorough understanding of the distribution and prevalence of these approaches within the medical literature. Figure 7 illustrates the distribution of cost-sensitive approaches used in the selected studies.

Direct approaches account for the largest share of papers, representing 76% (133 papers) of the qualified studies, indicating a clear focus on integrating cost-sensitive considerations directly into the learning process. Some researchers modified the objective function of the model to minimise the expected cost of misclassification. For example, Al-Sawwa and Ludwig (2019) introduced a new objective function within their cost-sensitive centroid-based differential evolution classification algorithm. This function involves two key steps: allocating misclassification costs to each class label and evaluating the fitness of individual vectors from the population. Initially, instances are assigned to the closest centroid based on the Euclidean distance. Subsequently, the misclassification cost is computed by summing over the misclassified instances. Other works incorporated the cost matrix directly into the loss function. For instance, in (Naceur et al. 2020), the authors used a weighted cross-entropy loss function in their Convolutional Neural Network (CNN) model for brain tumor segmentation. Furthermore, researchers have explored the fusion of multiple loss functions to tackle diverse problems effectively. Notably, one study fused the Focal loss with the Dice loss (Wang et al. 2022), while another investigation integrated the Dice loss with the weighted cross-entropy (Taghanaki et al. 2019). The ease of implementation is the primary factor contributing to this trend since most ML libraries offer readily available

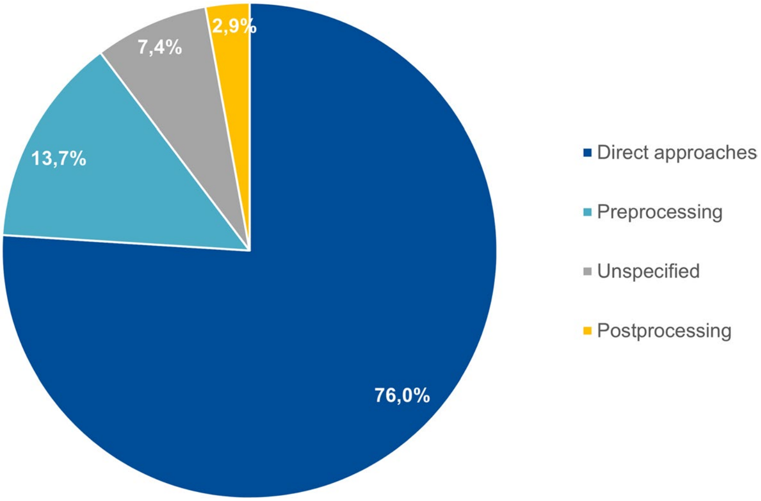

Fig. 7 Distribution of the selected studies per CSL approach

implementations (Sterner et al. 2021). Moreover, certain packages provide the flexibility to apply custom loss functions directly to the algorithm, allowing users to employ cost-sensitive loss functions tailored to their specific applications.

A considerable share of the selected studies (16.6%) adopted meta-learning approaches. Precisely, preprocessing was applied in 24 papers (13.7%), and postprocessing was employed in 5 papers (2.9%). Preprocessing was carried out using weighting or MetaCost. For instance, Wang et al. (2013) implemented cost-sensitive logistic regression and decision trees by weighting training instances based on the total cost associated with each class in the provided cost matrix. In another study, Afzal et al. (2013) employed MetaCost to integrate cost sensitivity into four ML models. Preprocessing techniques are adopted by researchers as they alter the training data instead of the underlying algorithm (Fernández et al. 2018), rendering them a suitable approach for different types of classifiers. On the other hand, postprocessing relied on thresholding. In a study conducted by Zhao et al. (2018), empirical thresholding was employed to iteratively adjust the decision threshold, aiming to select the classifier with minimal total cost. Alternatively, Liu et al. (2021) determined the threshold as per Eq. 3, leveraging prior knowledge about misclassification costs when developing a multi-label ECG classifier. This method demonstrated superior performance compared to commonly used thresholding techniques, including rank-based thresholding, proportion-based thresholding, and fixed thresholding. Thresholding is less frequently used in the selected studies due to the computational challenge of tuning multiple thresholds (equivalent to the number of labels considered), particularly in multi-label classification (Liu et al. 2021).

Note that direct and preprocessing approaches were utilised together in two papers, resulting in double counting in these categories. Additionally, 13 articles (7.4%) did not provide information on the cost-sensitive approach they adopted and were thus categorised as "unspecified". Incomplete reporting may hinder the reproducibility and comparability of results and the identification of effective methods for dealing with imbalanced medical data. Given the importance of transparency in medical research, future studies should provide a clear and detailed description of the implemented cost-sensitive techniques, including any modifications made to the model, to allow for better understanding, comparison and replication of findings.

6 Strengths and weaknesses

In this section, a detailed examination of the strengths and weaknesses of CSL is conducted across different dimensions. The section is divided into three subsections: the first explores the general strengths and limitations of CSL, the second delves into the strengths and limitations of CSL approaches, and the third analyses the reported strengths and limitations in selected works.

6.1 Strengths and weaknesses of CSL

CSL, as applied to imbalanced medical data, presents a unique framework that brings both advantages and limitations to the table. Understanding these strengths and weaknesses is crucial to making informed decisions and driving advancements in the field. Dedicated tables have been prepared to provide a comprehensive overview of these aspects.

Table 12 presents the strengths and weaknesses of CSL in addressing class-imbalanced medical datasets. CSL techniques offer a multitude of advantages when dealing with such challenging scenarios. Firstly, they efficiently mitigate class imbalance, leading to more balanced predictions and improved performance across all classes. Notably, an overwhelming majority of the selected studies (97.7%, 169 papers) reported enhanced performance compared to cost-insensitive methods, state-of-the-art models, or other balancing techniques. Moreover, CSL explicitly considers the unequal misclassification costs in cost-sensitive problems, particularly inherent in medical decision-making (Liu et al. 2021; Siddiqui et al. 2020; Wang and Cheng 2021), ensuring that the model's predictions align with the critical consequences of FP and FN. Importantly, these techniques achieve these benefits without altering the underlying data distribution, conserving the integrity and representativeness of the dataset. This preservation of the original data structure allows for full utilisation of all available data, in contrast to resampling techniques. Additionally, CSL techniques exhibit computational efficiency, enabling their application to large-scale medical datasets without excessive resource requirements. This finding aligns with the broader consensus from other reviews (Haixiang et al. 2017; Kaur et al. 2019; Tarekegn et al. 2021). Lastly, they prove particularly effective in handling severely class-imbalanced scenarios where conventional learning algorithms struggle to provide accurate predictions. This notable characteristic has also been highlighted in the survey conducted by Leevy et al. (2018), reaffirming the effectiveness of CSL in addressing highly imbalanced data. Together, these advantages position CSL as a valuable tool in tackling class imbalance and enhancing the reliability and applicability of ML models in challenging medical contexts.

Building upon the strengths and benefits of CSL, it is essential to also acknowledge the associated limitations and challenges that require careful consideration. One significant concern arises from the unknown nature of misclassification cost values, a challenge echoed in previous review studies (He and Garcia 2009; Haixiang et al. 2017; Kaur et al. 2019; Tarekegn et al. 2021; Leevy et al. 2018; Johnson and Khoshgoftaar 2019; Elrahman and Abraham 2013; Sun et al. 2011; Feng et al. 2020). Accurately defining the costs of misclassifying different classes can be intricate and challenging. The design of the cost matrix often requires expert judgment and domain-specific knowledge (Fernando and Tsokos

Table 12 Strengths and weaknesses of CSL

Strengths	Studies	Weaknesses	Studies
Mitigates the class imbalance efficiently	Naceur et al. (2020); Zubair and Yoon (2022); Siddiqui et al. (2020); Chanchal et al. (2022); Shan et al. (2023); Qian et al. (2022); Jiang et al. (2017); Ebiaredoh-Mienye et al. (2022); Shen et al. (2020); Munagala et al. (2022); Newaz et al. (2021); Ravi (2022)	Misclassification cost values are unknown	Aldraimli et al. (2021); Liu et al. (2021); Naceur et al. (2020); Afzal et al. (2013); Zhao et al. (2018); Siddiqui et al. (2020); Fernando and Tsokos (2022); Nunes et al. (2013); Naceur et al. (2019); Kumar and Thakur (2021); Zhao et al. (2022); Cao et al. (2013b); Lili et al. (2016); Ravi et al. (2022)
Takes into account the unequal misclassification costs in cost-sensitive problems	Uguroglu et al. (2012); Wang et al. (2018a); Pranto et al. (2020); Gan et al. (2020); Fan et al. (2022); Liu et al. (2021); Wang and Cheng (2021); Zhenya and Zhang (2021); Ormeño et al. (2012); Jiang and Zhao (2021); Daraei and Hamidi (2017); Barot and Jethva (2021b)	Risk of overfitting the under-represented classes	Zhao et al. (2022); Li et al. (2022a)
Effective when dealing with severely class-imbalanced scenarios	Galdran et al. (2020); Siddiqui et al. (2020); Lee et al. (2023); Wang et al. (2020b)		
Does not alter the data distribution	Fan et al. (2022); Yang et al. 2021); Wang et al. (2013); Zubair and Yoon (2022); Sadeghi et al. (2022); Wang and Cheng (2021); Nunes et al. (2013); Jiang et al. (2017); Raj et al. (2021); Chamseddine et al. (2022); Mienye and Sun (2021); Zeng et al. (2021); Sheng et al. (2021); Cazañas-Gordón et al. (2022); Castro et al. (2020)		
Computationally efficient	Lee et al. (2023); Gour and Khanna (2022)		

2022), making it a delicate and complex task, especially considering that such expertise and knowledge may not always be readily available or accessible.

One strategy to address the issue of unknown costs is to conduct a thorough search for the optimal cost setup (Nunes et al. 2013). This approach entails exploring various combinations or configurations of cost values and assessing their impact on the performance of the cost-sensitive model. Techniques such as cross-validation or grid search can be employed to identify the best cost setup by iteratively testing and evaluating different cost values using predefined performance metrics. Additionally, the literature proposes specific cost assignment strategies. For example, a potential solution suggested in previous research (Haixiang et al. 2017) consists of setting the misclassification cost of the majority class to 1 and equating the penalty for the minority class to the IR. This approach has been adopted by several selected studies (e.g., (Fan et al. 2022; Wang et al. 2013; Liu et al. 2019; Wang et al. 2020c; Ashfaq et al. 2019; Hashemi et al. 2018), while other studies have proposed alternative cost formulas (e.g., (Zieba et al. 2014; Calderon-Ramirez et al. 2021; Roy et al. 2022; Yao et al. 2022; Zieba 2014).

In a noteworthy contribution, a study (Gan et al. 2020) highlighted a particular concern about the prevalent use of fixed misclassification costs in most CSL methods. In response to this constraint, researchers have investigated dynamic weight assignments during the training process. For instance, Focal loss (e.g., (Galdran et al. 2020; Li et al. 2022b; Lu et al. 2021; Naseem et al. 2020; Shirokikh et al. 2020; Lee et al. 2023)) adapts the costs based on the varying difficulty or significance of individual samples. Another study (Liu et al. 2019) incorporates an online-learning step to dynamically reweight each batch of the training set based on its validation performance.

Another limitation that warrants attention is the risk of overfitting the under-represented classes. When misclassification costs are inadequately defined and heavily weighted towards the minority classes, CSL methods can exhibit excessive adaptation to these classes (Sun et al. 2011), which may lead to overfitting (Elrahman and Abraham 2013) and reduced generalization performance. Therefore, thoughtful attention should be paid to defining the costs, ensuring their appropriateness, and mitigating the risks associated with excessive adaptation.

6.2 Strengths and weaknesses of CSL approaches

Transitioning to the analysis of CSL approaches, it is essential to acknowledge their inheritance of the broader advantages and disadvantages of CSL. Additionally, they exhibit their own specific strengths and weaknesses, which are succinctly outlined in Table 13 for reference. It should be noted that some of these particular strengths and weaknesses are derived from existing reviews beyond the scope of the selected papers in this study.

Direct approaches offer both advantages and disadvantages. On the positive side, these approaches benefit from the availability of readily implemented solutions in many ML libraries. This accessibility allows researchers and practitioners to apply CSL techniques easily without extensive coding efforts. However, it is important to note that direct modifications in the learning algorithm, such as modifying the Gini index to account for misclassification costs in decision trees (Barot and Jethva 2021b, a), require a deep understanding of the underlying algorithms. This requirement means that researchers and practitioners must possess comprehensive knowledge of the specific algorithms being utilised. Another limitation of direct approaches is their potentially reduced versatility compared to other CSL approaches. By directly modifying the learning algorithm, these approaches are often

Table 13 Strengths and weaknesses of CSL approaches

Approach	Strengths	Studies	Weaknesses	Studies
Direct	Many available ML libraries	Sterner et al. (2021)	Require a deep understanding of the underlying learning algorithms	Sterner et al. (2021; Sun et al. (2011)
			Reduced versatility	Liu et al. (2021)
Preprocessing (Weighting)	Simple	Kaur et al. (2019); Sun et al. (2011); Qin et al. (2010)	–	–
	Flexible	Kaur et al. (2019)		
	Do not modify the learning algorithm	López et al. (2013); Fernández et al. (2018); Ling and Sheng (2008); Wang et al. (2013); Afzal et al. (2013); Sun et al. (2011)		
Preprocessing (MetaCost)	Flexible	Fernández et al. (2018)	Additional computational steps during the training phase	Afzal et al. (2013); Rekha et al. (2019)
	Do not modify the learning algorithm	López et al. (2013); Fernández et al. (2018); Ling and Sheng (2008); Wang et al. (2013); Afzal et al. (2013)		
Postprocessing (Thresholding)	Flexible	Liu et al. (2021)	Creating a division between training and cost-sensitive evaluation	Fernández et al. (2018)
	Do not modify the learning algorithm	López et al. (2013); Fernández et al. (2018); Ling and Sheng (2008); Wang et al. (2013); Afzal et al. (2013); Feng et al. (2020)	The number of thresholds that need to be tuned is usually no less than the number of considered labels	Liu et al. (2021)

customised for specific ML models, constraining their applicability across a broader range of models.

Preprocessing approaches, particularly weighting, offer several advantages in the context of CSL. Firstly, weighting is a simple technique that is easy to implement and interpret, as it involves adjusting the weights assigned to training instances. Moreover, it exhibits high flexibility by not necessitating any alterations to the underlying learning algorithm, ensuring adaptability to various ML models. As such, weighting emerges as a versatile approach for incorporating cost sensitivity.

Similarly, one key strength of MetaCost as a preprocessing approach is its flexibility, which allows for adapting classifiers to cost-sensitive scenarios without altering the underlying learning algorithm, operating at the instance level. This preserves compatibility with different ML models. Nevertheless, it is crucial to consider the computational implications of MetaCost, as it introduces additional steps during the training phase, such as data relabelling, which may result in increased computational complexity and longer training times.

In the realm of postprocessing approaches, particularly thresholding, a set of distinct advantages and limitations emerges. Notably, thresholding techniques exhibit a remarkable level of flexibility, enabling the adjustment of the classification model to various cost definitions without the need for retraining (Liu et al. 2021). This flexibility empowers researchers to adaptably fine-tune the model's behaviour to align with specific cost-sensitive requirements. Furthermore, it is worth highlighting that thresholding does not require modifications to the underlying learning algorithm. However, it is important to acknowledge the limitations associated with thresholding techniques. One such limitation lies in the division between the training and the subsequent cost-sensitive evaluation phases. During the initial training, where cost information is unavailable, the classifier is driven by error minimisation rather than cost optimisation (Fernández et al. 2018). This implies that the estimation of cost parameters is initialised using a cost-insensitive method, which may introduce inherent biases into the outcomes. The problem is effectively addressed by incorporating an ROC-based criterion into classifier training, as performance for both classes can be evaluated at once (Fernández et al. 2018). Another challenge is the tuning of thresholds, as the number of thresholds that need to be adjusted is typically no less than the number of considered labels. This process can be time-consuming and may require careful fine-tuning to achieve optimal cost-sensitive performance.

Expanding upon our exploration of the strengths and weaknesses of CSL approaches, it is crucial to consider the valuable insights derived from other studies in the field. Recent research (Vanderschueren et al. 2022) categorises CSL approaches based on the stage at which misclassification costs are incorporated into two classes: cost-sensitive training of models and cost-sensitive decision-making. The former category encompasses techniques applied before or during model training to construct a classifier, notably direct approaches and meta-learning preprocessing methods. Conversely, the latter category concerns thresholding techniques used after training to inform decision-making processes. Regarding performance, models utilising thresholding may yield superior overall predictive accuracy; however, those adopting cost-sensitive training excel in their ability to make high-quality decisions, focusing on accurate predictions as they impact decision outcomes. Notably, the study, conducted on nine datasets from various application areas, revealed that training a cost-insensitive model and subsequently introducing misclassification costs during the test phase through thresholding can be conceptually straightforward and effective. Nevertheless, it was highlighted that, under specific conditions, cost-sensitive training may emerge as the optimal choice. For instance, in cases of model misspecification, adopting a cost-sensitive objective function may outperform thresholding. Furthermore, the investigation indicated that combining cost-sensitive training and thresholding may not consistently enhance performance. In light of these results, we hold the view that exploring the combination of CSL approaches presents a promising avenue that merits further experimentation and investigation.

In a previous comparison of instance weighting and thresholding (Zhao 2008), it was observed that instance weighting is computationally more demanding, particularly in scenarios with uncertain cost settings. Unlike thresholding, which adjusts the decision threshold, instance weighting requires retraining the classifier when changing cost settings. An interesting observation from this study was that when instance weighting is applied and the classifier is insensitive to the cost ratio, the thresholding technique suffices. However, when the classifier is highly sensitive to the cost ratio under instance weighting, it becomes crucial to incorporate misclassification costs during training. These findings underscore the importance of conducting more extensive evaluations and comparisons of CSL approaches to validate these observations, as we emphasise the need for further research in this direction.

Building on these insights, it becomes evident that the choice of a CSL approach in medical applications, especially in cases characterised by imbalanced class distributions, has significant implications for model performance. Researchers and practitioners should commence by evaluating the computational resources at their disposal. Preprocessing techniques such as weighting offer an expedient and versatile avenue for implementation, requiring no modifications to the underlying algorithms. However, direct approaches may offer tailored solutions for individuals with a deep understanding of algorithms, emphasising the need to balance computational efficiency and model customisation. Moreover, the tuning task when dealing with thresholding necessitates anticipation. Additionally, staying abreast of the most recent advancements in CSL is crucial, as new techniques may yield enhanced results. Regardless of the chosen CSL approach, thorough validation on the specific medical dataset at hand remains non-negotiable. This validation ensures the alignment of the selected approach with the data's inherent characteristics and the fulfilment of particular research objectives. Consequently, well-informed decisions can be made to enhance model performance across various medical disciplines and tasks.

6.3 Strengths and weaknesses highlighted in certain selected works

Examining the implemented methods in the selected research necessitates a thorough assessment of their strengths and weaknesses. Table 14 compares various proposed methods to facilitate this evaluation, encompassing their main tasks, data types, employed CSL techniques, weighting formulas, and the reported advantages and limitations for each method.

The information provided in Table 14 illuminates several key trends and findings. Firstly, it is evident that the proposed solutions encompass a wide range of models, spanning from traditional ML to DL architectures. This diversity highlights the versatility and adaptability of CSL across different modelling paradigms. Furthermore, the application of CSL is not limited to specific types of data. The selected studies showcase the use of CSL for various data types, including numerical data, categorical data, images, time series, and textual data. This broad utilisation of CSL underscores its effectiveness in addressing cost-sensitive challenges across diverse data modalities.

A common thread among the presented works is the consistent achievement of enhanced performance by leveraging CSL techniques, effectively addressing the challenges posed by

Table 14 Strengths and weaknesses of proposed cost-sensitive methods in the selected studies

Study	Task	Data type	Proposed method	CSL technique	Cost values	Advantages	Limitations
Zhenya and Zhang (2021)	Heart disease diagnosis	Numeric, Categorical	Weighted Voting Ensemble (Random Forest, Logistic regression, SVM, ELM and KNN)	Weighting individual classifiers	Determined based on financial costs	Better results compared to single classifiers and previous studies The limitations of a particular classifier are remedied by other classifiers Good generalization ability Closer to reality by considering misclassification costs	Longer training time State-of-the-art techniques such as DL and soft computing, which would improve the performance, are not included
Barot and Jethva (2021a)	Breast cancer diagnosis	Numeric, Categorical	Decision trees	Integrating costs into Gini index calculation	-	Better results compared to previous studies Reduces misclassification costs More balanced performance for both classes	-
Razzaghi et al. (2016)	Patient financial risk prediction Patient vaccination prediction following a reminder	-	Multilevel SVM	Weighting the regularization parameter C	Selected as inversely proportional to the size of each class	Improved results Reduced computational time Robust	-
Liu et al. (2018)	Breast cancer diagnosis	Numeric	SVM	Weighting the regularization parameter C	Quantified based on misclassification consequences	Improved performance Increased specificity	Decreased sensitivity (but a better overall performance)

Table 14 (continued)

Study	Task	Data type	Proposed method	CSL technique	Cost values	Advantages	Limitations
Uguroglu et al. (2012)	Heart disease diagnosis	Numeric, Categorical, Time series (ECG records)	KNN	Weighting the neighbours' votes	$C_n = 1$ and $C_p = \frac{K}{2} + 1$ where K is the number of neighbours	Better results compared to state-of-the-art methods and other ML models Can apply to a larger population High AUC scores using the least invasive, least costly and least risky tests	-
Lee et al. (2023)	Sleep stage classification	Time series (ECG records)	A DL architecture integrating a CNN, a Bidirectional Long Short-Term Memory (Bi-LSTM), and a Soft Voting-based Ensemble	Focal loss	-	Better performance than state-of-the-art methods Solves the class imbalance problem Reduces the training time Avoids overfitting	May not work well for subjects with sleep disorders that have different sleep structures than healthy subjects No external validation was performed Potential delays if applied in online and real-time applications
Holste et al. (2022)	Thorax disease diagnosis	Images (Chest X-ray)	A deep CNN-based model (ResNet50)	Weighted loss functions: weighted Cross-Entropy (CE) loss, Focal loss, Label-Distribution-Aware Margin loss, Influence-Balanced loss	-	Improved performance for infrequent classes	Which re-weighting method provides more significant gains appears to depend on its interaction with the loss function used
Ashfaq et al. (2019)	Hospital readmission prediction of congestive heart failure patients	Numeric, Categorical	LSTM	Weighted CE loss	Determined based on the IR	Enhanced performance Fast training time Increased sensitivity Cost savings	Train and test data are from a single region

Table 14 (continued)

Study	Task	Data type	Proposed method	CSL technique	Cost values	Advantages	Limitations
Wang et al. (2018a)	Hospital readmission prediction	Time series (Numeric, Categorical)	A DL model integrating a CNN and a multilayer perception	Weighted CE loss	Determined based on the IR	Better results compared to state-of-the-art models Deployed in a real system for readmission prediction	Moderate sensitivity
Fernando and Tsokos (2022)	Skin lesion diagnosis	Images (Dermoscopy images)	A deep CNN-based model (EfficientNet)	Dynamically weighted balanced loss function (composed of two terms: dynamically weighted CE and a regularization component equal to the entropy of Brier score)	$C_i = log(\frac{n_m}{n_i}) + 1$ where n_m is the frequency of the majority class, and n_i is the frequency of class i	Better results compared to CE loss, weighted CE loss, and Focal loss Dynamic weighting (self-adapt its weights depending on the prediction scores) with an emphasis on hard-to-train examples Robust generalization Broad applicability (medicine and intrusion detection applications)	-
Javidi et al. (2021)	COVID-19 early detection	Images (CT scans)	A deep CNN-based model (hybrid DenseNet and CapsNet)	Weighted loss function	$C_p = 1 - \frac{n_p}{N}$ and $C_n = 1 - \frac{n_n}{N}$ where n_p and n_n are the frequencies of the positive and negatives classes, and N is the total number of samples	Improved performance Robust (even if the positive samples are 50 times less than the negative samples) Stable Fast convergence Can process large images even with a small number of training data	The output does not contain any explicit segmentation of diagnostically helpful components

Table 14 (continued)

Study	Task	Data type	Proposed method	CSL technique	Cost values	Advantages	Limitations
Liu et al. (2021)	Cardiovascular diseases screening	Time series (ECG records)	A deep CNN-based model (ResNet)	Thresholding	Determined by domain experts	Better results compared to other commonly used thresholding methods (rank-based thresholding, proportion-based thresholding, and fixed thresholding) Ranked among the top 10 teams in the PhysioNet/CinC challenge	Potential loss of cost information (the cost information in the cost matrix is converted to the costs for binary classification) Unreasonable predictions cannot be avoided for this multi-label ECG classification task (two labels can be predicted to coexist in a recording) Lack of interpretability
Zhang and Shen (2011)	Alzheimer's disease diagnosis	Images (MRI, PET), Numeric	SVM	Thresholding	Fixed costs	Improved performance Multi-stage cost-sensitive model (integrating cost-sensitivity at the feature selection and classification stages)	-
Zhao et al. (2018)	Medical incidents detection due to look-alike sound-alike (LASA) mix-ups	Text	Logistic regression	Thresholding	Testing multiple values	Improved performance	Outperformed by resampling (perhaps due to the uncertainty and inconsistency of the cost matrix in training and testing the dataset)

Table 14 (continued)

Study	Task	Data type	Proposed method	CSL technique	Cost values	Advantages	Limitations
Cao et al. (2013a)	Lung nodule detection	Images (CT scans)	Adaptive Random Subspace Ensemble	Thresholding	Determined using a heuristic search strategy with G-mean as the fitness function	Better results compared to resampling and Adacost Improved generalization	-
Reychav et al. (2019)	Cardiac patient survival prediction in emergency situations	Numeric, Categorical	Logistic regression	Thresholding	Testing multiple values	Improved performance	The optimal proportions of positive and negative samples in the training data were not computed when dividing the data into train/test The data (Israel) might not be generalizable to other areas Needs to be further validated and tested with other datasets
Shen et al. (2022)	Sleep apnea detection	Times series (PPG signals)	A DL model integrating a deep CNN (multi-attention ResNet) and AdaCost	Weighting	Determined based on the IR	Better results compared to state-of-the-art models Effectively reduces the gap between specificity and sensitivity Lower running time Real-time detection	Sensitivity could be further enhanced Complex structure and numerous parameters
Hsu et al. (2015)	Breast cancer risk assessment	Numeric, Categorical	Decision trees, Logistic model tree, Naïve Bayes, SVM, KNN, Radial basis function network	Weighting	Testing multiple values	Better performance than sampling and ensemble learning Perfect recall score (100%)	Reasonable precision

Table 14 (continued)

Study	Task	Data type	Proposed method	CSL technique	Cost values	Advantages	Limitations
Li et al. (2021)	Brain tumour classification Lung cancer staging	Images (MRI, CT scans)	A 3D Siamese network (self-supervised)	Weighting	$C_p = \frac{N}{n_p}$ and $C_n = \frac{N}{n_n}$	Better results A large boost in predicting the minor class Successfully tackles class imbalance	-
Henze et al. (2021)	Seizure detection	Time series (ECG records)	Naïve Bayes, KNN, SVM, Adaboost	Weighting	Determined based on the IR	High sensitivity Lower detection latency	High false alarm rate Data quality (improvements in the technical setting might lead to better availability and quality of the heart rate data)
Zhang et al. (2018)	Breast cancer diagnosis Liver disease diagnosis	Numeric, Categorical	Hierarchical ELM	Weighting	Determined based on the IR	Effectively solve the class imbalance problem with small biomedical datasets Higher and more stable performance than other state-of-the-art methods Enhanced generalization	-
Wang et al. (2013)	Survivability prognosis of breast cancer	Numeric, Categorical	Logistic regression, Decision trees	Weighting	Determined based on the IR	Better results compared to resampling and ensemble learning Higher predictive performance	-

Table 14 (continued)

Study	Task	Data type	Proposed method	CSL technique	Cost values	Advantages	Limitations
Sung et al. (2021)	Acute stroke diagnosis	Numeric, Categorical	Random forest, SVM, Logistic regression, Decision trees, KNN	Weighting	-	Enhanced performance	Outperformed by resampling Generalizability needs further examination (single-site study) Prehospital factors (such as mode of transportation to the hospital and diagnosis by emergency medical services) were not considered

class-imbalanced medical datasets. Moreover, many methods exhibit reduced computational time indicating their efficiency and practicality in real-world applications. However, one work (Zhenya and Zhang 2021) employing ensemble learning deviated from this trend, possibly due to the inherent complexity of the ensemble approach. Additionally, the robustness of the proposed methods is apparent, with several studies highlighting their ability to deliver reliable and stable results. This robustness enhances the trustworthiness of the proposed solutions in practical healthcare settings. Notably, one study (Ashfaq et al. 2019) even emphasised the potential cost savings associated with CSL, further underscoring the practical benefits of these approaches.

The papers also consistently underscore the significance of validation and generalizability in medical applications. Acknowledging the importance of validating the proposed models with diverse datasets and across different clinical sites reflects a commitment to ensuring the reliability and applicability of the methods in real-world scenarios. Furthermore, interpretability surfaces as a critical consideration in developing cost-sensitive solutions. One study (Liu et al. 2021) highlighted the challenge of interpretability as a potential limitation, recognising the need to balance model complexity and interpretability to facilitate transparency and understanding in clinical decision-making processes. Moreover, the studies shed light on the challenges related to the specificity and sensitivity trade-off, a common concern in CSL. Achieving an optimal balance between these measures is crucial for attaining accurate predictions while minimising false alarms.

It is also noteworthy that several works compared CSL with alternative strategies such as resampling and ensemble learning. The outcomes of these comparisons varied, with some studies showcasing the superior performance of CSL, while others found alternative strategies to be more effective. These findings highlight the importance of carefully selecting the most suitable strategy based on the specific characteristics of the dataset and the learning task at hand.

7 Datasets and data types

This section provides an overview of the datasets and data types utilised in the selected studies, shedding light on the variety and characteristics of the data employed to evaluate cost-sensitive methods.

7.1 Datasets

The medical datasets used in the selected studies are imbalanced and thus perfectly suitable to assess the performance of the developed cost-sensitive methods and evaluate their effectiveness. A total of 196 datasets were identified across the 173 selected papers. Note that 52 papers (30%) employed multiple datasets to evaluate their methods. Table 15 presents the most common datasets used in at least four studies, along with their sources, data types, number of instances, attributes, classes and papers. All the presented datasets are publicly available. The MIT-BIH Arrhythmia database was the most commonly used in seven selected studies, followed by COVID-19 Chest X-ray, Wisconsin Diagnostic Breast Cancer, and Pima Indians Diabetes datasets, each employed in six papers. The ISIC 2019 dataset was used in five papers, while Thyroid Disease, HAM10000, ILPD, and BUPA

Table 15 Most frequently used datasets in the selected studies

Dataset	Source	Data type	#Instances	#Attributes	#Classes	#Papers
MIT-BIH Arrhythmia	Moody and Mark (1980)	Time series (ECG records)	48	-	5	7
COVID-19 Chest X-ray	Cohen et al. (2020a)	Images (Chest X-ray), Numeric and Categorical (metadata)	123	16	5	6
Wisconsin (Diagnostic)	Wolberg et al. (1995)	Numeric	569	30	2	6
Pima Indians Diabetes	National Institute of Diabetes and Digestive and Kidney Diseases (1990)	Numeric	768	8	2	6
ISIC 2019	ISIC Challenge (2019)	Images (Dermoscopy)	25331	-	8	5
Thyroid Disease	Quinlan (1987)	Numeric, Categorical	3772	21	3	4
HAM10000	Tschandl (2018)	Images (Dermoscopy), Numeric and Categorical (metadata)	10015	5	7	4
Indian Liver Patient (ILPD)	Ramana and Venkateswarlu (2012)	Numeric, Categorical	583	10	2	4
BUPA Liver Disorders	UCI Machine Learning Repository (1990)	Numeric, Categorical	345	5	2	4

Liver Disorders datasets were each used in four papers. Of the 187 remaining datasets, 70% were publicly available, and 30% were acquired from hospitals and healthcare units.

Based on the findings, the MIT-BIH Arrhythmia Database (Moody and Mark 1980) was the most commonly used dataset, owing to the numerous selected studies in cardiology and its reputation as a well-known benchmark dataset. The MIT-BIH Arrhythmia Database comprises 48 ECG recordings, each with a duration of 30 min and a sampling frequency of 360 Hz (Shi et al. 2019). All heartbeats within the recordings are expertly annotated and categorised into one of 15 heartbeat types. The AAMI standard provides a standardised approach for labelling these arrhythmias to ensure consistency and comparability across different studies. According to this standard, heartbeats are recommended to be grouped into five main classes (Shi et al. 2019): Normal (N), Supraventricular ectopic beat (S), Ventricular ectopic beat (V), Fusion of ventricular and normal beat (F), and Unknown beat (Q). Although most studies adhere to this standard (Li et al. 2022b; Zhao et al. 2023; Wang et al. 2019; Han et al. 2022), some may choose alternative classification schemes (Lu et al. 2021). The extensive use of this database has resulted in numerous published studies, providing ample opportunities for comparisons with previous results. The reported IRs of the database differ based on the class distribution employed for the classification task. One such study, conducted by Zubair and Yoon (2022), reported an IR of approximately 9:1, with normal beats accounting for 89.5% of the dataset and abnormal beats accounting for the remaining 10.5%. This severe class imbalance makes the MIT-BIH Arrhythmia Database an exceedingly challenging dataset for cost-insensitive models.

The COVID-19 Chest X-ray dataset (Cohen et al. 2020a), assembled by Cohen and colleagues (2020b) in February 2020 from publicly available sources, has gained widespread recognition as a reference dataset for developing and evaluating DL algorithms to detect COVID-19 from chest X-ray images. Comprising five distinct types of pneumonia cases, including COVID-19, SARS, Streptococcus spp., Pneumocystis spp., and ARDS, this dataset has been extensively utilised as a starting point for exploring various DL techniques, particularly during the surge of research amidst the COVID-19 pandemic. Its public availability and recognition have made it a convenient choice for researchers to compare their findings with other studies in the field. Moreover, the dataset's acknowledged class imbalance, with a higher number of COVID-19-positive cases compared to other respiratory diseases, makes it a relevant and challenging testbed for evaluating the performance of CSL algorithms. Consequently, the COVID-19 Chest X-ray dataset is frequently employed in the selected studies for these compelling reasons. Nevertheless, it is noteworthy that the dataset has certain limitations, as it may not provide a fully comprehensive or representative sample of the general population. This is likely why the six studies that employed the dataset incorporated additional datasets to supplement their findings. In September 2020, the dataset was updated with 679 frontal chest X-ray images from 412 individuals in 26 countries (Cohen et al. 2020c).

The Wisconsin Diagnostic Breast Cancer dataset (Wolberg et al. 1995) has been extensively employed in numerous studies, primarily due to the prominence of oncology in the selected research. Of the 54 studies conducted in this sub-field, 38.9% focused on breast cancer, making it a prevalent research topic. Moreover, the dataset's imbalanced distribution, with a disproportionate number of malignant cases (212) compared to benign cases (357), makes it a relevant resource for research on CSL in the medical field. The Pima Indians Diabetes dataset (National Institute of Diabetes and Digestive and Kidney Diseases 1990) has garnered equal attention in the selected research papers owing to its pertinence in the diagnosis of diabetes, a critical healthcare concern that affects a large population worldwide. Moreover, the dataset suffers from a significant class imbalance between the presence (34.9%) and absence (65.1%) of diabetes cases, rendering it an appropriate choice for evaluating the performance of cost-sensitive models.

In the selected research, using multiple datasets to evaluate model performance was commonplace. This approach is highly beneficial as it enables the assessment of model generalizability, facilitates a more comprehensive evaluation, and establishes benchmarks for the field. Datasets with varying difficulty levels, IRs, and feature spaces can expose the strengths and limitations of models. Using multiple datasets mitigates the risk of models overfitting to one particular dataset, ensuring models are more readily applicable to new and unseen data.

7.2 Data types

Data represents a fundamental aspect of CSL in the medical field, and comprehending the used data types is paramount. The analysis of the selected papers revealed a diverse range of attribute types. Of the 173 studies analysed, 78 papers (45.1%) utilised images as the primary data type, while numerical and categorical data were combined in 51 papers (29.5%). Additionally, time series (24 papers, 13.9%), textual data (seven papers, 4%), and numeric data (four papers, 2.3%) were also employed. Less prevalent combinations included images with numerical and categorical data (seven papers, 4%), time series with text (one paper, 0.6%), time series with numerical data (one paper, 0.6%), time series with numerical and categorical data (one paper, 0.6%), and images with numerical data (one paper, 0.6%). For a more granular perspective, the most prominent types of medical imagery data included Computed Tomography (CT) scans (19 papers), Magnetic Resonance Imaging (MRI) (16 papers), Chest X-ray (13 papers), dermoscopic images (10 papers), fundus photographs (eight papers) and thermograms (four papers). Regarding time series data, the most commonly used types were Electrocardiogram (ECG) records (nine papers), Electroencephalogram (EEG) records (five papers), and Cardiotocography (CTG) records (two papers).

Regardless of the data types, medical data is inherently prone to imbalance due to the nature of medical conditions and patient populations. The analysis revealed a broad spectrum of data types among the selected studies. This diversity in data types can be justified by the fact that different medical applications require different types of data for accurate outcomes. For instance, medical imaging techniques, such as CT scans, MRI, and Chest X-rays, are essential for detecting anomalies in anatomical structures. In contrast, time-series data, such as ECG and EEG records, are necessary to monitor physiological function changes over time.

The dominance of images as the primary data type can be attributed to the increasing use of medical imaging techniques in clinical practice and research. With the advancements in imaging technology, clinicians can now obtain high-quality images that provide detailed information about the structure and function of various organs and tissues. Additionally, numerical data can capture continuous measurements such as blood pressure, heart rate, and body temperature. In contrast, categorical data can capture non-continuous measurements such as gender, age, and medical history. Combining these two data types can help achieve a more comprehensive understanding of a patient's health status. Moreover, the use of time series data highlights the importance of temporal information in medical applications. Time series data can capture changes in a patient's health status throughout time, aiding in the detection and prediction of medical conditions.

Among the selected research, the presence of text data was also observed. Textual data can capture unstructured information like clinical notes, medical reports, and patient history. Such data can provide valuable insights into the subjective nature of a patient's medical condition, including their symptoms, emotions, and experiences, which may not be accurately captured by numerical or categorical data alone. Additionally, textual data can be leveraged to identify patterns and relationships between medical conditions and patient characteristics, paving the way for developing personalised treatment plans.

8 Performance metrics

Assessing the performance of cost-sensitive algorithms in medical applications requires appropriate performance metrics. Two categories of metrics commonly used in the literature are traditional metrics and cost-related metrics. This study focuses on presenting the commonly used metrics from both categories.

8.1 Traditional metrics

Traditional performance metrics, such as accuracy, precision, sensitivity, and F1 score, provide insights into the overall predictive performance of cost-sensitive algorithms. These metrics evaluate the model's ability to correctly classify instances without explicitly considering the cost associated with misclassifications. To quantify these metrics, a fundamental tool called a confusion matrix is employed.

The confusion matrix, showcased in Table 16, provides a detailed breakdown of the model's predictions and the actual class labels. It summarises the counts of True Positives (TP), True Negatives (TN), FP, and FN. TP corresponds to the accurately predicted positive instances, while TN represents the accurately predicted negative instances. Conversely, FP and FN denote instances that were erroneously classified as positive or negative, respectively.

- Accuracy is a standard evaluation measure in ML used to assess a model's ability to predict class labels accurately. It is defined as the ratio of correct predictions to the total number of predictions made:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

  Relying solely on accuracy may not be appropriate for imbalanced datasets, as it can cause misleading results, where a model that appears to perform well may, in fact, be biased towards the majority class. To avoid such bias, all the selected studies using accuracy, except one (Naseem et al. 2020), utilised complementary metrics to evaluate model performance comprehensively.
- Error rate is the complement of accuracy. It quantifies the percentage of misclassified instances and is calculated as follows:

Table 16 Confusion matrix

	Actual negative	Actual positive
Predicted negative	TN	FN
Predicted positive	FP	TP

$$Error\ rate = 1 - Accuracy = \frac{FP + FN}{TP + FP + TN + FN} \tag{10}$$

- Sensitivity, also called recall or True Positive Rate (TPR), quantifies the proportion of TP predictions among all positive predictions:

$$Sensitivity = \frac{TP}{TP + FN} \tag{11}$$

  A high sensitivity value in medical contexts is particularly desirable as it reduces the risk of FN and ensures the correct identification of all positive cases. This holds crucial significance in medical diagnosis, where detecting all individuals with the disease (TP) is paramount, and missing a diagnosis can lead to delayed treatment and severe health complications.
- Specificity, also known as the True Negative Rate (TNR), measures the proportion of TN predictions relative to all negative predictions:

$$Specificity = \frac{TN}{TN + FP} \tag{12}$$

  In medical settings, a high specificity rating is critical to reduce the occurrence of FP and ensure accurate identification of all negative cases. FP can lead to unnecessary medical interventions or additional diagnostic procedures, highlighting the critical role of specificity in reliable medical diagnosis.
- Precision is a performance metric that quantifies the accuracy of positive predictions made by a model. It is computed as follows:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

  It is worth noting that precision and sensitivity demonstrate an inverse correlation, whereby improving one metric often leads to a decline in the other. When dealing with imbalanced medical data, prioritising sensitivity at the expense of precision may result in increased FP, resulting in unwarranted medical interventions or additional tests. Thus, precision becomes a crucial metric in evaluating the performance of a model that seeks to minimise the number of FP while maximising TP.
- The AUC metric quantifies a model's ability to discern between positive and negative cases, rendering it a compelling choice for medical applications. The AUC metric ranges between 0 and 1, with a higher value indicating better overall performance. The AUC is computed as the area under the Receiver Operating Characteristic (ROC) curve, which graphically represents the model's TPR plotted against the False Positive Rate (FPR) at varying threshold levels. The FPR can be derived as the complement of specificity:

$$FPR = 1 - Specificity \tag{14}$$

  The ROC curve visually illustrates the trade-off between sensitivity and specificity across different classification thresholds. Frequently paired with the AUC metric, the ROC curve facilitates visual comparison and evaluation of diverse models' performances. In medical research and decision-making, the ROC curve and AUC metric assume significance by aiding in selecting an optimal threshold that strikes a balance between sensitivity and specificity, catering to the specific requirements of the medical task at hand. An ideal classifier would be positioned in the top-left corner, representing a perfect balance between sensitivity and specificity. The closer the ROC curve of a model approaches this ideal point, the better its performance.

- The Geometric Mean (G-Mean) is another performance metric that comprehensively evaluates a model's accuracy by combining sensitivity and specificity. It is commonly employed in scenarios involving imbalanced datasets. G-mean is defined as the geometric mean of sensitivity and specificity:

$$G\text{-}mean = \sqrt{Sensitivity \cdot Specificity} \tag{15}$$

By considering sensitivity and specificity, the G-Mean offers a balanced assessment of a classifier's performance on both minority and majority classes. It provides a reliable measure of accuracy, accounting for the occurrence of FP and FN. This metric is particularly valuable in medical datasets where the costs associated with FP and FN can vary significantly.

- The balanced accuracy metric offers a comprehensive evaluation of a classifier's accuracy on both positive and negative classes, taking into account both sensitivity and specificity. It is calculated as the average of sensitivity and specificity:

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} \tag{16}$$

Balanced accuracy provides an equitable assessment by considering the performance on both classes equally, addressing the potential bias towards the majority class in the traditional accuracy measure. This makes it particularly suitable for evaluating classifiers in scenarios with imbalanced datasets and enhances its clinical relevance as an evaluation criterion.

- F1 score, also known as the F-measure, combines precision and sensitivity to assess the overall effectiveness of a classifier. It provides a balanced evaluation by considering the model's ability to correctly identify positive instances (precision) and capture all positive instances (sensitivity). The F1 score is calculated as the harmonic mean of precision and sensitivity, ensuring that both measures are considered equally:

$$F1\ score = \frac{2 \cdot (Precision \cdot Sensitivity)}{Precision + Sensitivity} \tag{17}$$

The F1 score finds particular utility in scenarios where both precision and sensitivity are important, such as medical diagnosis.

Moreover, the F-measure encompasses a range of metrics beyond the F1 score. These metrics, collectively called F$_\beta$ scores, introduce a parameter β that allows for flexible weighting of precision and sensitivity based on specific application requirements. The F$_\beta$ score is calculated using the following formula:

$$F_\beta\ score = \frac{(1+\beta^2) \cdot (Precision \cdot Sensitivity)}{(\beta^2 \cdot Precision) + Sensitivity} \tag{18}$$

The β parameter controls the relative emphasis placed on precision versus sensitivity. A higher β value (e.g., F2 score) favours sensitivity over precision, making it suitable when the cost of FN is significant. Conversely, a lower β value (e.g., F0.5 score) emphasises precision, making it appropriate when the cost of FP is more critical.

- The Area Under the Precision-Recall Curve (AUPRC) serves as a comprehensive measure of a classifier's overall effectiveness in capturing positive instances across different classification thresholds. In contrast to the ROC curve, which considers the trade-off between sensitivity and specificity, the Precision-Recall (PR) curve focuses on the trade-off between precision and sensitivity (recall). The PR curve plots precision values against corresponding sensitivity values at various thresholds.

  The AUPRC is computed as the area under the PR curve. Spanning the interval of 0 to 1, a higher AUPRC value reflects superior performance, indicating that the classifier achieves high precision while maintaining a high sensitivity rate. This implies that the classifier accurately identifies positive instances while minimising FP. The AUPRC metric is especially beneficial with datasets exhibiting significant class imbalance or when the consequences of FN and FP differ, as commonly seen in medical applications.
- The Matthews Correlation Coefficient (MCC) metric measures the quality of binary classifiers, taking into account TP, TN, FP, and FN. MCC is calculated using the following formula:

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{19}$$

  MCC ranges from -1 to +1, where a score of 1 indicates a perfect prediction, 0 represents a random prediction, and -1 indicates a complete disagreement between the prediction and the actual label.

  MCC is commonly used in fields such as bioinformatics, where imbalanced datasets and binary classification problems are prevalent (Chicco and Jurman 2020). It is considered a robust statistical measure (Sadeghi et al. 2022) as it yields a high score only when the predictions exhibit strong performance across all four categories of the confusion matrix (TP, FP, TN, and FN).
- The Kappa score, also known as Cohen's Kappa, is a statistical measure that assesses the level of agreement between two annotators or raters in categorical classification tasks. It considers both the accuracy of the classifier and the possibility of agreement occurring by chance. The Kappa score is computed via the subsequent formula, where p_0 is the observed agreement or accuracy (the proportion of instances where the classifier and the actual labels agree) and p_e is the expected agreement (the agreement expected by chance alone):

$$Kappa = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e} \tag{20}$$

  p_e is calculated based on the marginal probabilities of the classifier's predictions and the true labels.

  The Kappa score ranges from -1 to 1, with higher values indicating a higher level of agreement between the classifier's predictions and the true labels. A score of 1 represents a perfect agreement beyond chance, 0 indicates agreement equivalent to chance, and negative values indicate less agreement than expected by chance.

  The Kappa score is instrumental in situations with a class imbalance or where relying solely on accuracy can be misleading. It serves as a useful metric for assessing the consistency and reliability of categorical classifications, providing insights into the quality of annotations or the performance of classifiers compared to human annotators.

8.2 Cost-related metrics

Cost-related metrics thoroughly evaluate classifier performance by explicitly incorporating the costs associated with different misclassifications. These metrics go beyond traditional metrics and take into account the real-world impact of classification errors. By considering the consequences of misclassification, cost-related metrics offer a more nuanced and practical assessment of a classifier's effectiveness, especially in domains where misclassification costs are high, such as medicine.

- The Misclassification Cost (MC) metric, sometimes called average cost (Guido et al. 2022), provides a comprehensive evaluation of a classifier's performance by considering the potential costs associated with misclassifying instances in a classification task. Unlike traditional accuracy, which treats all misclassifications equally, the MC metric assigns specific costs to different types of errors based on their impact or significance in a given application. This performance measure enables a more informed evaluation in cost-sensitive applications, as is the case for medical ones. The MC metric is calculated as follows:

$$MC = \frac{(FP \cdot C_p) + (FN \cdot C_n)}{TP + TN + FP + FN} \tag{21}$$

- The cost curve is a graphical depiction representing a binary classifier's performance (expected cost) over the full range of possible class distributions and misclassification costs (Drummond and Holte 2000, 2006). The y-axis corresponds to the normalised expected cost, which can be computed using the following formula:

$$EC_{\text{norm}} = \frac{FN \cdot P(+) \cdot C_p + FP \cdot (1 - P(+)) \cdot C_n}{P(+) \cdot C_p + (1 - P(+)) \cdot C_n} \tag{22}$$

  where $P(+)$ is the probability of an example being from the positive class.

  The x-axis corresponds to the "probability times cost", which summarises misclassification costs and class distributions in a single number:

$$P(+) \cdot \text{cost} = \frac{P(+) \cdot C_p}{P(+) \cdot C_p + (1 - P(+)) \cdot C_n} \tag{23}$$

  Drummond and Holte (2000) introduced cost curves as a remedy to address the limitations of ROC curves. Cost curves offer a comprehensive evaluation of classifier performance by considering specific misclassification costs, class probabilities, performance comparisons between different classifiers, average performance across multiple evaluations, confidence intervals, and statistical significance of performance differences, making them a powerful tool for decision-making in classification tasks.

  In their notable research, Drummond and Holte (2006) provide an illustrative example that complements their prior work. The illustration showcases the cost lines associated with C4.5 decision trees and 1R models on the Japanese credit dataset, where costs are taken into consideration.

- The weighted Kappa score (Cohen 1968) is a modified version of the Kappa score, which incorporates weights (the misclassification costs) that reflect the severity or importance of disagreement for each class based on a cost matrix. It can be formulated as follows:

$$Kappa_w = \frac{\sum_{i=1}^{I}\sum_{j=1}^{I} w_{ij}\cdot P_{ij} - \sum_{i=1}^{I}\sum_{j=1}^{I} w_{ij}\cdot P_{i.}P_{.j}}{1-\sum_{i=1}^{I}\sum_{j=1}^{I} w_{ij}\cdot P_{i.}P_{.j}} \tag{24}$$

where w_{ij} refers to the weight associated with the value in the ith row and jth column of the confusion matrix, and $P_{i.}$ and $P_{.j}$ are the marginal probabilities.

Weighted Kappa is a valuable tool for effectively addressing cost-sensitive classification (Ben-David 2008). In situations where misclassification costs are unknown and can only be estimated, conducting sensitivity analysis using weighted Kappa is strongly advised (Ben-David 2008).

- Cost-Weighted Accuracy (CWA), proposed by the PhysioNet/CinC challenge (Alday et al. 2020), is a multi-class scoring metric that extends the traditional accuracy metric by incorporating cost-based weights. To calculate the cost-weighted accuracy, the prediction results are organised in a multi-class confusion matrix denoted by $A = [a_{ij}]$, where a_{ij} represents the number of instances from class j classified as class i. The scoring is derived by performing a weighted averaging of the matrix A, where each entry is multiplied by its corresponding cost-based weight, w_{ij}:

$$CWA = \sum_{i,j} a_{ij}w_{ij} \tag{25}$$

Here, w_{ij} represents the cost of misclassifying an instance of class j into class i based on treatment similarities or differences in risks. The score is then normalised to range between 0 and 1, where a perfect classifier receives a score of 1 for correctly predicting the true labels, and an inactive classifier gets a score of 0 for always predicting the normal class. The scoring metric fully acknowledges and rewards accurate diagnoses while granting partial credit to misdiagnoses with similar risks or outcomes as the true diagnosis.

8.3 The distribution of performance metrics in the selected studies

Figure 8 presents the most commonly used performance metrics in the selected studies. Among these metrics, sensitivity was the most frequently used, appearing in 139 papers. Accuracy and specificity followed, being utilised in 100 and 91 studies, respectively. AUC and precision were employed in 77 papers each, while the ROC curve was utilised in 54 papers. G-mean was observed in 43 papers. The Dice score and balanced accuracy exhibited similar usage, being used in 14 studies each. On the other hand, FPR and MC were employed in nine studies each. Less common metrics included, among others, AUPRC and FNR, which were utilised in five and four papers, respectively.

In addition to the MC metric, other cost-related metrics were also relatively underutilised. Specifically, the CWA metric was identified in only one paper, while cost curves and the weighted Kappa score were not employed. This finding highlights the need for increased emphasis on incorporating and exploring cost-related metrics in CSL research.

It is also noteworthy that the vast majority of the selected studies incorporated several metrics in their performance evaluation. Combining multiple evaluation metrics is a widely adopted practice in ML research, particularly when dealing with imbalanced datasets. This approach ensures a comprehensive assessment of the model's performance. Additionally, different metrics can capture distinct aspects of model performance. By integrating multiple metrics, researchers can obtain a more nuanced understanding of the model's strengths

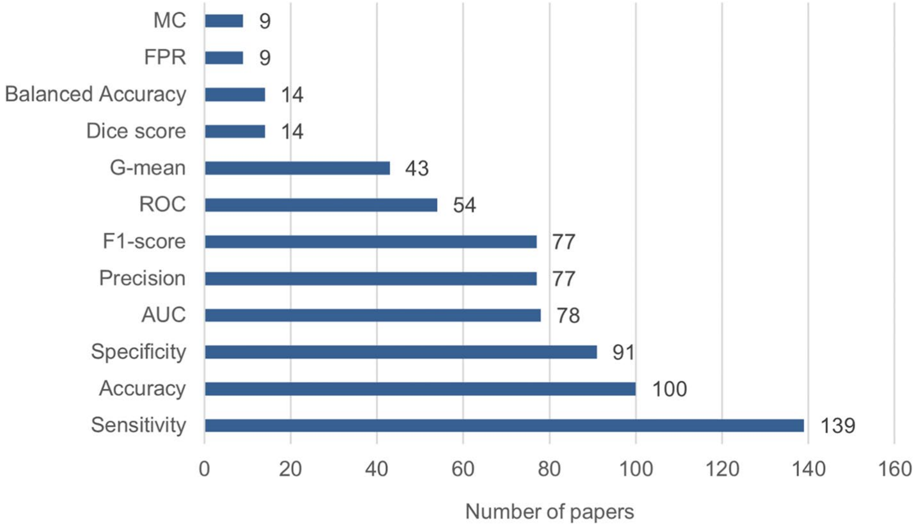

Fig. 8 Most commonly used metrics in the selected studies

and limitations, allowing them to make informed decisions about which model best suits a given problem.

9 Development tools

Identifying the development tools employed in implementing CSL techniques is crucial for understanding the technical landscape of this research domain. This information provides insights into the practical aspects and technical capabilities of CSL models, thereby facilitating reproducibility, collaboration, and future advancements in the field. However, a significant challenge arose as a considerable number of papers (63, 36.4%) did not disclose the specific tools adopted. The ones employed in the remaining studies were ascertained through explicit disclosure or meticulous analysis of shared code. Notably, our investigation focuses on the tools used for implementing the cost-sensitive property within CSL techniques, distinct from those employed for statistical analysis, preprocessing, plotting, and other related tasks. The identified tools, along with their corresponding frequencies and license information, are presented in Table 17.

Among the selected studies, Python (2023), an open-source programming language, emerged as the most widely used tool, appearing in 64 papers (37%). Python's extensive usage can be attributed to its popularity in the ML community, its ease of use for prototyping and experimentation, and its rich ecosystem of libraries and frameworks. Notably, libraries such as TensorFlow (2023b), Keras (2023), PyTorch (2023a), Scikit-learn (2023a), XGBoost (2022), and LightGBM (2023) have contributed significantly to Python's prominence and were widely employed in the selected studies. These libraries offer comprehensive support for implementing ML models, providing various functionalities, including handling class imbalance.

Table 17 Tools used in the selected studies for CSL techniques' implementation

Tool	License	#Papers
Python	Open-source	64
Weka	Open-source	14
MATLAB	Proprietary	15
R	Open-source	9
Libsvm	Open-source	7
KEEL	Open-source	2
Caffe	Open-source	2
Java	Open-source	1
RapidMiner	Commercial	1

TensorFlow and Keras provide options to implement the cost-sensitive property by weighting the loss function, enabling researchers to assign higher importance to minority classes and effectively address class imbalance. One example of such functionality is the `weighted_cross_entropy_with_logits` function (TensorFlow 2023a) in TensorFlow, which allows applying class weights directly to the loss calculation using the `pos_weight` parameter. Furthermore, Keras offers the `class_weight` parameter in its models, enabling users to set class weights during training, thereby enhancing the capability to handle class imbalance within the Keras framework. Scikit-learn also provides the `class_weight` parameter in various classifiers. Additionally, XGBoost, a popular gradient boosting library in Python, offers the `scale_pos_weight` parameter. By assigning a higher weight to the minority class, XGBoost ensures balanced learning and improved performance on imbalanced datasets. LightGBM, another robust gradient boosting library in Python, enables handling class imbalance through the `class_weight` parameter. By assigning appropriate weights to different classes, LightGBM adjusts the impact of each class during model training, leading to enhanced performance on imbalanced data.

PyTorch is another popular library widely used for implementing DL models. While PyTorch does not provide a specific parameter or functionality for directly handling class imbalance in the same way as the aforementioned libraries, it offers a flexible and customisable framework that allows researchers to implement various techniques for CSL. In PyTorch, researchers can manually assign class weights during the training process by modifying the loss function. By multiplying the loss for each sample by its corresponding class weight, researchers can give higher importance to the minority classes. This approach allows for fine-grained control and customisation in handling class imbalance based on the specific requirements of the problem. Additionally, PyTorch integrates well with other Python libraries, such as Scikit-learn, which offers class weight support in its classifiers. Researchers have the option to utilise the `compute_class_weight` (Scikit-learn 2023b) function provided by Scikit-learn to calculate weights, which can then be incorporated into the weight parameter of the `CrossEntropyLoss` function (PyTorch 2023b), for example.

MATLAB (2023b), a popular proprietary programming environment and language, was the second most frequently employed, featuring in 15 papers (8.7%). MATLAB's widespread adoption in various scientific domains, including ML and data analysis, explains its presence in the selected research. Its extensive collection of toolboxes offers a wide range of pre-built algorithms and functions that facilitate implementing ML models, including those designed for cost-sensitive applications. MATLAB supports CSL through two key parameters in the fitting functions: `Cost`, (MATLAB 2023a) which utilises cost matrices to represent misclassification costs for different classes, and ClassWeights, (MATLAB 2023c) allowing users to assign specific weights to each class during training. These features provide flexibility in addressing class imbalance and implementing effective cost-sensitive models. Moreover, MATLAB's efficient handling of large datasets, high-performance computing capabilities, compatibility with other programming languages, and active user community contribute to its popularity and usability in ML and data analysis. These additional advantages, combined with its support for CSL, solidify MATLAB's position as a versatile and powerful tool for implementing cost-sensitive models.

WEKA (2023), an open-source tool, ranked as the third most utilised tool, accounting for 14 papers (8.1%). Its inclusion can be attributed to its comprehensive set of ML tools and algorithms, its support for CSL, and its large user base. Weka offers users the choice between a user-friendly Graphical User Interface (GUI) and a Java API, catering to individuals with different preferences and programming expertise. The GUI serves as an accessible option for users with limited programming knowledge, while the Java API provides greater control and flexibility for those with programming skills. In the context of CSL, Weka provides wrappers and meta-classifiers that streamline the implementation process. One such meta-classifier is the `CostSensitiveClassifier` (Trigg 2023a), which enables users to transform a base classifier in Weka into a cost-sensitive model by incorporating a cost matrix during model training. The cost matrix can be conveniently specified as input, facilitating the automatic handling of class imbalance. Another available meta-classifier is `MetaCost` (Trigg 2023b).

The open-source programming language R (2023) appeared prominently in 9 papers (5.2%) within the selected studies, owing to its extensive collection of packages designed for ML and statistical modelling. R offers a wide range of packages that support CSL. For instance, the `Mlr` package enables users to implement cost-sensitive models through thresholding and weighting techniques (Bischl et al. 2022). Another package, `Caret`, provides a unified interface for training and evaluating various ML models, offering CSL support through the `train` function (Kuhn 2008). This function allows users to specify the misclassification cost for each class using the `weights` argument. Similarly, the `Rpart` library incorporates the `weights` argument (R 2022b), allowing users to assign different weights to classes while constructing decision trees. Additionally, the `LiblineaR` package permits the assignment of higher weights to instances of the minority class using the `wi` argument during the development of linear models (R 2022a). Moreover, the active community support, flexibility, and seamless integration with complementary data manipulation and visualisation tools further contribute to R's prominence in CSL research.

Despite appearing in fewer papers (7 papers, 4%), LibSVM's inclusion in the selected studies highlights its notable contributions to CSL research. LibSVM (Chang and Lin 2023) is an open-source software package renowned for its efficiency and flexibility in implementing SVM algorithms. One of LibSVM's key strengths lies in its ability to support CSL. This is achieved by leveraging the `-wi` option, which allows specific weight values to be assigned to each class during model training. Furthermore, LibSVM's multi-language support and extensive documentation further enhance its recognition and adoption in the research community, solidifying its position as a valuable tool for researchers exploring CSL.

KEEL (2018), an open-source Java software tool, appeared twice (1.2%) in the selected research. KEEL is specifically designed for knowledge data discovery tasks. Its user-friendly interface and wide range of functionalities make it a valuable resource for ML researchers and educators. KEEL provides pre-built, ready-to-use cost-sensitive versions of popular ML algorithms, namely C4.5 decision trees (Ting 2002), multilayer perception (Zhou and Liu 2006), SVM (Tang et al. 2009), and Adaptive Boosting (AdaBoost) (Sun et al. 2007).

The open-source DL framework Caffe (Jia et al. 2014) and the programming language Java (Oracle 2023) had low occurrences in the selected studies. Caffe was utilised in two papers (1.2%), while Java appeared only in one paper (0.6%). These low occurrences can be attributed to multiple factors. One reason is that, as per our understanding, Caffe and Java do not offer built-in functionalities for CSL. However, researchers have the flexibility to implement cost-sensitive models by developing custom functions tailored to their specific needs. Additionally, it is worth mentioning that the merger of Caffe2 into PyTorch in 2018 (Caffe2 2018) may have contributed to the underutilisation of Caffe, as researchers have increasingly migrated to PyTorch for its comprehensive support and capabilities.

Likewise, the commercial data science platform RapidMiner (2023a) made a modest appearance in the selected research, with only one paper (0.6%) acknowledging its presence. To implement CSL, users can utilise the software's cost-sensitive operator (RapidMiner 2023b), conveniently accessible through its GUI. This operator allows specifying the costs associated with different classes and incorporates these costs into the learning process. In addition, RapidMiner offers a dedicated implementation of the MetaCost algorithm through the MetaCost operator (RapidMiner 2023c).

A noteworthy observation in this study is that some papers combined two development tools, resulting in their double counting in both respective tool categories. Among them, three papers (1.7%) utilised both LibSVM and MATLAB (Liu et al. 2018; Razzaghi et al. 2015; Prashanth and Roy 2018), while one paper (0.6%) employed Python alongside MATLAB (Rahman et al. 2021a). Additionally, one study (0.6%) leveraged the combined capabilities of Python and Weka (Wu et al. 2020). These combinations highlight researchers' versatility and adaptability in harnessing various tools to address their specific research objectives.

10 Limitations

This section aims to critically examine the limitations encountered in this study, which are primarily associated with the following factors:

- Selection bias: Various measures were taken to minimise potential selection bias in this review. A comprehensive search strategy was implemented, incorporating a diverse set of search terms, alternative spellings, and synonyms. The search comprised all article fields and was carried out across multiple databases, including PubMed, IEEE Xplore, Springer Link, Science Direct, and Google Scholar. The inclusion of Google Scholar was explicitly intended to retrieve papers that may not have been available in the first four libraries. Moreover, the selection criteria were rigorously defined and applied carefully to the candidate papers by one author while the remaining authors evaluated the final selection. Any disagreements between the three authors were resolved through meetings until a consensus was reached. To reduce exclusions, reasonable QA criteria were designed to ensure that papers of sufficient quality were included in the study. Besides, theoretical papers and reviews were assessed using only two non-empirical QA questions to avoid overlooking them. Despite these efforts, some limitations should be acknowledged. It is plausible that some relevant works may have been missed, specifically those published in languages other than English, in other databases, or in non-peer-reviewed sources not encompassed in the search. Additionally, snowballing (i.e. manual search in reference lists) was not conducted, which could have identified additional relevant studies.

- Data extraction bias: In light of the critical and time-intensive nature of data extraction, a meticulous approach was taken to mitigate potential bias. One author conducted the task carefully, while the other two authors diligently reviewed the extracted data to ensure its accuracy and impartiality. Despite our efforts, some degree of subjectivity may have been introduced. Regular meetings were held to reconcile divergences and achieve a mutually agreed-upon interpretation of the data to counteract this possibility.

11 Implications for future research

This section discusses the key implications of our review and provides practical guidance for researchers. We outline the following implications that can drive advancements in this field and support practical applications:

- Understanding domain-specific imbalance: Imbalanced medical datasets present substantial challenges arising from the inherent characteristics of medical data, where certain conditions exhibit significantly lower prevalence than others. To address these challenges, researchers must cultivate a profound understanding of the specific class imbalance issues within their targeted medical domain. This necessitates a comprehensive examination of the distribution patterns of medical conditions in the dataset, the identification of critical minority classes, and an exploration of the underlying factors contributing to this imbalance. Furthermore, researchers must evaluate the potential consequences of misclassification within their specific medical context. This evaluation entails a thorough consideration of the associated risks, costs, and implications associated with FN and FP.
- Cost matrix design and evaluation: As CSL relies on the accurate estimation of misclassification costs, researchers should carefully consider the design and evaluation of the cost matrix. Collaborating with domain experts and healthcare professionals is highly valuable to define the costs associated with different types of misclassifications, especially in medical settings where the consequences of FN and FP can differ significantly. Researchers are encouraged to explore methods for cost matrix estimation, including expert opinions, data-driven approaches, and incorporating contextual factors.
- Combining attribute and misclassification costs: While misclassification costs capture the consequences of FN and FP, attribute costs reflect the challenges associated with acquiring specific features, encompassing aspects such as financial expenses, time constraints, or the invasiveness of required tests (Fernández et al. 2018). By combining these two types of costs, researchers can develop comprehensive cost-sensitive models that simultaneously account for predictive performance and cost-efficiency in feature selection. Striking an optimal balance between the performance achieved by utilising certain features and the costs associated with their acquisition enables the development of more effective and resource-efficient models for medical decision-making.
- Hybrid CSL: Combining CSL with other balancing strategies presents a promising avenue for addressing class imbalance in medical datasets. By integrating CSL with strategies like resampling or ensemble learning, researchers can leverage the strengths of multiple strategies to handle class imbalance and address the associated

misclassification costs effectively. It is crucial, however, to gain a deep understanding of the characteristics and requirements of the specific medical dataset under investigation. This understanding allows for identifying scenarios where CSL or other balancing strategies excel individually and situations where combining them yields better results, ultimately leading to more effective and tailored solutions for imbalanced medical data.

- Cost-sensitive evaluation: Traditional performance metrics may not fully capture the effectiveness of models when misclassification costs are unequally distributed. Researchers are strongly encouraged to expand the evaluation beyond conventional metrics and employ cost-sensitive metrics that directly incorporate the associated misclassification costs. Additionally, a comprehensive evaluation strategy should combine multiple metrics to gain a holistic understanding of the model's performance in terms of both classification accuracy and cost-effectiveness.
- Addressing less investigated medical disciplines and tasks: While disciplines such as oncology, cardiology, neurology, and infectious diseases have garnered significant research attention, other medical disciplines have received relatively less investigation. Similarly, diagnosis has been extensively studied, while other medical tasks remain relatively unexplored. To address this gap, researchers are urged to broaden their focus beyond the well-investigated medical sub-fields and tasks and delve into the untapped potential of less investigated medical domains. This exploration will facilitate a deeper understanding of the applicability and effectiveness of CSL in a broader range of medical applications. Furthermore, promoting data sharing is highly recommended, as limited dataset availability may have contributed to the underrepresentation of certain medical disciplines and tasks in the existing literature.
- Advancing validation research: The scarcity of papers dedicated to validation research reflects the inherent challenges in conducting assessments of cost-sensitive methods in real-world hospital settings. Therefore, researchers must establish close collaborations with medical professionals and actively engage in validation studies to demonstrate the effectiveness and reliability of CSL methods in real medical scenarios. These validation studies can provide valuable insights into the practical performance of CSL models and enhance the trust and confidence of healthcare practitioners.
- Ensuring generalizability: Researchers should focus on developing models that can effectively handle class imbalance across diverse datasets and healthcare contexts. This involves evaluating the performance of cost-sensitive methods on multiple datasets, encompassing different medical institutions and patient populations. Furthermore, efforts should be made to address potential sources of dataset bias, covariate shift, and concept drift to enhance the models' generalizability to unseen data.
- Considering interpretability: Interpretability is recognised as a critical consideration in developing cost-sensitive solutions. Guaranteeing interpretability within these models is paramount for cultivating transparency and understanding in clinical decision-making processes. Researchers are urged to prioritise the development of interpretable cost-sensitive techniques that strike a balance between model complexity and transparency. This emphasis empowers medical professionals to accurately interpret and trust the predictions made by these models.
- Enhancing reproducibility: Researchers are encouraged to actively engage in data and code-sharing practices, fostering a collaborative environment that enables the scientific community to reproduce and validate research findings. Furthermore, it is crucial to provide detailed reports on the cost-sensitive methods employed, covering the specific cost matrix used, the chosen cost-sensitive approach, and the algorithmic configura-

tions. Comprehensive reporting promotes transparency and facilitates comparisons between different CSL methods.

- Broader applicability: Researchers should extend their focus beyond single-label classification and segmentation tasks and dedicate more attention to multi-label (Tarekegn et al. 2021) and regression (Wang et al. 2020a) problems. While single-label classification and segmentation have received significant attention, there is a need for comprehensive investigations and advancements in cost-sensitive methods for tasks involving multiple labels and continuous outcome prediction. By broadening the scope of CSL to encompass these diverse problems, researchers can expand the applicability of CSL in a wider range of medical scenarios.

12 Conclusion and future work

This review aimed to provide an overview of the available literature on CSL for imbalanced medical data. Our study marks a novel contribution to the domain, being the first of its kind. A total of 173 papers published between January 2010 and December 2022 and sourced from five digital libraries were carefully selected, analysed and classified.

The results demonstrated an apparent rise in interest and research activity in CSL for imbalanced medical data, particularly since 2020. This growing recognition underscores the challenges of class imbalance in medical datasets and the pressing need for effective solutions. The substantial number of papers published in renowned journals further highlights the scholarly significance and impact of this research area. Among the selected works, a considerable portion focused on proposing novel solutions and evaluating their effectiveness, demonstrating the dual nature of research efforts in addressing class imbalance. The prevalence of HBE as the primary empirical type suggests the extensive use of past data to assess the performance of CSL methods. This practice is driven by the abundance of available historical datasets and the challenges associated with accessing real-world medical data.

Furthermore, the investigation of medical sub-fields revealed that oncology received the highest level of attention, emphasising the critical importance of accurate prediction and diagnosis in cancer-related applications. In parallel, the prominence of diagnosis as the most widely studied medical task underscores the significance of precise and timely diagnostic capabilities in medical decision-making. Notably, researchers have displayed a strong preference for CSL direct approaches, highlighting the relevance of integrating cost sensitivity directly into the learning process. This preference may be attributed to the availability of readily implemented solutions in popular ML libraries. This study also detailedly explored the strengths and weaknesses of CSL strategy and approaches, equipping researchers with crucial insights and recommendations to make informed decisions. In addition, a comparative analysis was conducted on a selection of relevant works, allowing for a deeper understanding of the performance and characteristics of different CSL techniques.

For datasets, a total of 196 datasets were identified from the selected papers. The common practice of using multiple datasets enabled the assessment of model generalizability and facilitated a more comprehensive evaluation of cost-sensitive methods. The findings additionally showed that MIT-BIH Arrhythmia emerged as the most frequently used dataset, owing to the numerous selected studies in cardiology and its recognition as a reputable benchmark dataset. Moreover, the analysis of data types revealed a wide range of modalities, with images being the dominant type. This observation aligns with the increasing utilisation of medical imaging techniques in clinical practice and research.

The evaluation metrics employed in the selected studies encompassed two categories: traditional metrics and cost-related metrics. The underutilisation of cost-related metrics stresses the need for their incorporation in CSL research. Additionally, combining multiple evaluation metrics was commonplace, ensuring a comprehensive assessment of model performance, particularly in the context of imbalanced datasets. This study also shed light on the development tools employed for CSL implementation. Nine tools were identified, each accompanied by a detailed explanation of how they incorporate CSL. Python emerged as the most widely adopted programming tool, aligning with its popularity in the broader ML community.

Lastly, this paper elucidated several significant implications, offering researchers valuable insights and practical guidance and contributing to the advancement of the field. First and foremost, understanding the specific nature of class imbalance within medical datasets is crucial for developing effective CSL techniques. Additionally, the design and evaluation of cost matrices play a pivotal role in accurately reflecting the misclassification costs associated with different classes. Furthermore, combining attribute costs with misclassification costs enables researchers to develop comprehensive cost-sensitive models that balance performance and cost efficiency in feature selection. Integrating CSL with other balancing techniques offers promising avenues for addressing class imbalance in medical datasets. Moreover, the study highlighted the importance of adopting cost-sensitive evaluation metrics that go beyond traditional performance measures to capture the true impact of misclassification costs. The study also emphasised the need to address less investigated medical disciplines and tasks and advance validation research to demonstrate the effectiveness and reliability of CSL models. Additionally, ensuring the generalizability of CSL techniques across different medical datasets and settings is crucial for their practical application and broader adoption in the field. Considering the interpretability of CSL models is also essential for fostering trust and transparency in medical decision-making processes. Furthermore, promoting reproducibility by sharing datasets, code, and detailed reporting of CSL approaches enhances collaboration and facilitates comparisons between different methods. Lastly, there is a need for further exploration and application of CSL techniques on multilabel and regression problems. Such efforts hold great potential for advancing accurate and cost-aware modelling in various imbalanced contexts.

We posit that our study will offer researchers and practitioners pertinent insights into the current landscape of CSL literature in medicine, along with recommendations for subsequent publications. Moreover, this study serves as a foundational step for our future research, which will entail a more focused and comprehensive systematic literature review on the performance evaluation of cost-sensitive techniques for imbalanced medical data. By narrowing down the research scope and delving deeper into specific performance aspects, our future work aims to provide a more nuanced understanding of the effectiveness and limitations of CSL in addressing class imbalance in the medical domain.

Acknowledgements This work was conducted under the research project "Machine Learning based Breast Cancer Diagnosis and Treatment", 2020-2023. We would like to thank the Moroccan Ministry of Higher Education and Scientific Research, the Digital Development Agency (ADD), the National Centre for Scientific and Technical Research (CNRST), and Mohammed VI Polytechnic University (UM6P) for their support. We also extend our gratitude to Google for generously supporting the first author of this paper through the Google PhD Fellowship.

Author contributions I.A. and A.I. conceived the research idea. I.A. performed data collection and analysis and drafted the manuscript. A.I. assisted in data analysis and critically revised the manuscript for intellectual content. I.C. provided a critical evaluation of the data analysis.

Declarations

Conflict of interest The authors declare no competing interests.

References

Afzal Z, Schuemie MJ, Blijderveen JCV et al (2013) Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. BMC Medical Informatics and Decision Making 13:1–11. https://doi.org/10.1186/1472-6947-13-30

Al-Sawwa J, Ludwig SA (2019) A cost-sensitive centroid-based differential evolution classification algorithm applied to cancer data sets. 2019 IEEE Symposium Series on Computational Intelligence. SSCI 2019:2514–2521. https://doi.org/10.1109/SSCI44817.2019.9002660

Alday EAP, Gu A, Shah AJ et al (2020) Classification of 12-lead ECGS: the physionet/computing in cardiology challenge 2020. Physiol Meas 41:124003. https://doi.org/10.1088/1361-6579/ABC960

Aldraimli M, Soria D, Grishchuck D et al (2021) A data science approach for early-stage prediction of patient's susceptibility to acute side effects of advanced radiotherapy. Comput Biol Med. https://doi.org/10.1016/J.COMPBIOMED.2021.104624

Aldraimli M, Osman S, Grishchuck D et al (2022) Development and optimization of a machine-learning prediction model for acute desquamation after breast radiation therapy in the multicenter requite cohort. Adv Radiat Oncol 7:100890. https://doi.org/10.1016/J.ADRO.2021.100890

Ashfaq A, Sant'Anna A, Lingman M et al (2019) Readmission prediction using deep learning on electronic health records. J Biomed Inf. https://doi.org/10.1016/J.JBI.2019.103256

Barot PA, Jethva HB (2021) Imbtree: minority class sensitive weighted decision tree for classification of unbalanced data. Int J Intell Syst Appl Eng 9:152–158. https://doi.org/10.18201/ijisae.2021473633

Barot PA, Jethva HB (2021) MGINI - improved decision tree using minority class sensitive splitting criterion for imbalanced data of Covid-19. J Inf Sci Eng 37:1097–1108. https://doi.org/10.6688/JISE.202109_37(5).0008

Ben-David A (2008) Comparison of classification accuracy using Cohen's weighted kappa. Exp Syst Appl 34:825–832. https://doi.org/10.1016/J.ESWA.2006.10.022

Bischl B, Lang M, Kotthoff L, et al (2022) Cost-sensitive classification mlr. Accessed 22 August 2023, https://mlr.mlr-org.com/articles/tutorial/cost_sensitive_classif.html

Breiman L, Friedman JH et al (1984) Classification and regression trees. Routledge. https://doi.org/10.1201/9781315139470

Caffe2 (2018) Caffe2 and PyTorch join forces to create a research + production platform. Accessed 22 August 2023, https://caffe2.ai/blog/2018/05/02/Caffe2_PyTorch_1_0.html

Calderon-Ramirez S, Yang S, Moemeni A et al (2021) Correcting data imbalance for semi-supervised Covid-19 detection using x-ray chest images. Appl Soft Comput. https://doi.org/10.1016/J.ASOC.2021.107692

Cao P, Zhao D, Zaiane O (2013a) Cost sensitive adaptive random subspace ensemble for computer-aided nodule detection. In: Proceedings of CBMS 2013 - 26th IEEE International Symposium on Computer-Based Medical Systems pp 173–178. https://doi.org/10.1109/CBMS.2013.6627784

Cao P, Zhao D, Zaiane O (2013) Measure oriented cost-sensitive SVM for 3D nodule detection. Ann Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Ann Int Conf 2013:3981–3984. https://doi.org/10.1109/EMBC.2013.6610417

Careers in medicine (2023) Specialty profiles. Accessed 22 August 2023, https://careersinmedicine.aamc.org/explore-options/specialty-profiles

Castro PB, Krohling B, Pacheco AG et al (2020) An app to detect melanoma using deep learning: an approach to handle imbalanced data based on evolutionary algorithms. Proc Int Joint Conf Neural Netw. https://doi.org/10.1109/IJCNN48605.2020.9207552

Cazañas-Gordón A, Parra-Mora E, Cruz LADS (2022) Distance-based loss weightings for improving retinal tissue segmentation using fully convolutional neural networks. In: 6th IEEE ecuador technical chapters meeting, ETCM 2022 https://doi.org/10.1109/ETCM56276.2022.9935708

Chamseddine E, Mansouri N, Soui M et al (2022) Handling class imbalance in Covid-19 chest x-ray images classification: using smote and weighted loss. Appl Soft Comput 129:109588. https://doi.org/10.1016/J.ASOC.2022.109588

Chanchal AK, Lal S, Kini J (2022) Deep structured residual encoder-decoder network with a novel loss function for nuclei segmentation of kidney and breast histopathology images. Multimed Tools Appl 81:9201–9224. https://doi.org/10.1007/S11042-021-11873-1

Chang CC, Lin CJ (2023) Libsvm a library for support vector machines. Accessed 22 August 2023, https://www.csie.ntu.edu.tw/~cjlin/libsvm/

Chicco D, Jurman G (2020) The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genom 21:1–13. https://doi.org/10.1186/S12864-019-6413-7

Cohen J (1968) Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 70:213–220. https://doi.org/10.1037/H0026256

Cohen JP, Morrison P, Dao L (2020a) Covid-19 chest x-ray database. Accessed 22 August 2023, https://github.com/ieee8023/covid-chestxray-dataset

Cohen JP, Morrison P, Dao L (2020b) Covid-19 image data collection. https://doi.org/10.48550/arxiv.2003.11597

Cohen JP, Morrison P, Dao L et al (2020) Covid-19 image data collection: prospective predictions are the future. Mach Learn Biomed Imag. https://doi.org/10.59275/j.melba.2020-48g7

Cunningham CT, Quan H, Hemmelgarn B et al (2015) Exploring physician specialist response rates to web-based surveys. BMC Med Res Methodol. https://doi.org/10.1186/S12874-015-0016-Z

Daraei A, Hamidi H (2017) An efficient predictive model for myocardial infarction using cost-sensitive j48 model. Iran J Publ Health 46:682

Devi D, Biswas SK, Purkayastha B (2019) A cost-sensitive weighted random forest technique for credit card fraud detection. In: 2019 10th international conference on computing, communication and networking technologies (ICCCNT), pp 1–6, https://doi.org/10.1109/ICCCNT45670.2019.8944885

Di Nunzio GM (2014) A new decision to take for cost-sensitive naïve bayes classifiers. Inf Process Manag 50(5):653–674. https://doi.org/10.1016/j.ipm.2014.04.008

Domingos P (1999) Metacost: A general method for making classifiers cost-sensitive. In: Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, NY, USA, KDD ’99, p 155-164, https://doi.org/10.1145/312129.312220

Dorado-Moreno M, Pérez-Ortiz M, Gutiérrez PA et al (2017) Dynamically weighted evolutionary ordinal neural network for solving an imbalanced liver transplantation problem. Artif Intell Med 77:1–11. https://doi.org/10.1016/J.ARTMED.2017.02.004

Drummond C, Holte RC (2000) Explicitly representing expected cost: an alternative to roc representation. In: Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, KDD ’00, pp198-207, https://doi.org/10.1145/347090.347126

Drummond C, Holte RC (2006) Cost curves: an improved method for visualizing classifier performance. Mach Learn 65:95–130. https://doi.org/10.1007/S10994-006-8199-5

Ebiaredoh-Mienye SA, Swart TG, Esenogho E (2022) A machine learning method with filter-based feature selection for improved prediction of chronic kidney disease. Bioengineering 9:350. https://doi.org/10.3390/BIOENGINEERING9080350

Elkan C (2001) The foundations of cost-sensitive learning. In: Proceedings of the 17th international joint conference on artificial intelligence - vol 2, IJCAI’01, pp973-978

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

Esfandiari N, Babavalian MR, Moghadam AME et al (2014) Knowledge discovery in medicine: current issue and future trend. Exp Syst Appl 41:4434–4463. https://doi.org/10.1016/J.ESWA.2014.01.011

Esteva A, Robicquet A, Ramsundar B et al (2019) A guide to deep learning in healthcare. Nat Med 25:24–29. https://doi.org/10.1038/s41591-018-0316-z

Fan B, Xie Z, Cheng H et al (2022) Risk prediction of diabetic readmission based on cost sensitive convolutional neural network. Commun Comput Inf Sci 1563:299–311. https://doi.org/10.1007/978-981-19-0852-1_23

Feng Y, Zhou M, Tong X (2020) Imbalanced classification: a paradigm-based review. Stat Anal Data Min 14:383–406. https://doi.org/10.1002/sam.11538

Fernandes K, Cardoso J, Fernandes J (2017) Cervical cancer (Risk Factors). UCI Mach Larn Repos https://doi.org/10.24432/C5Z310

Fernando KRM, Tsokos CP (2022) Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. IEEE Transact Neural Netw Learn Syst 33:2940–2951. https://doi.org/10.1109/TNNLS.2020.3047335

Fernández A, García S, Galar M et al (2018) Cost-sensitive learning. Learn Imbalanced Data Sets. https://doi.org/10.1007/978-3-319-98074-4_4

Freitas A, Brazdil P, Costa-Pereira A (2009) Cost-sensitive learning in medicine, IGI Global, pp 57–75. https://doi.org/10.4018/978-1-60566-218-3.ch003

Galar M, Fernandez A, Barrenechea E et al (2012) A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Transact Syst Man and Cybern Part C Appl Rev 42:463–484. https://doi.org/10.1109/TSMCC.2011.2161285

Galdran A, Dolz J, Chakor H, et al (2020) Cost-sensitive regularization for diabetic retinopathy grading from eye fundus images. Lecture notes in computer science (Including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 12265 LNCS:665–674. https://doi.org/10.1007/978-3-030-59722-1_64

Gan D, Shen J, An B et al (2020) Integrating tanbn with cost sensitive classification algorithm for imbalanced data in medical diagnosis. Comput Ind Eng 140:106266. https://doi.org/10.1016/J.CIE.2019.106266

Gour N, Khanna P (2022) Ocular diseases classification using a lightweight CNN and class weight balancing on oct images. Multimed Tools Appl 81:41765–41780. https://doi.org/10.1007/S11042-022-13617-1

Guido R, Groccia MC, Conforti D (2022) A hyper-parameter tuning approach for cost-sensitive support vector machine classifiers. Soft Comput. https://doi.org/10.1007/S00500-022-06768-8

Haixiang G, Yijing L, Shang J et al (2017) Learning from class-imbalanced data: review of methods and applications. Exp Syst Appl 73:220–239. https://doi.org/10.1016/J.ESWA.2016.12.035

Han C, Wang P, Huang R et al (2022) Hctnet: an experience-guided deep learning network for inter-patient arrhythmia classification on imbalanced dataset. Biomed Signal Process Control 78:103910. https://doi.org/10.1016/J.BSPC.2022.103910

Hashemi SR, Salehi SSM, Erdogmus D et al (2018) Asymmetric loss functions and deep densely connected networks for highly imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. IEEE Access 7:1721–1735. https://doi.org/10.1109/access.2018.2886371

He H, Garcia EA (2009) Learning from imbalanced data. IEEE Transact Knowl Data Eng 21:1263–1284. https://doi.org/10.1109/TKDE.2008.239

Henze J, Houta S, Surges R, et al (2021) Multimodal detection of tonic-clonic seizures based on 3d acceleration and heart rate data from an in-ear sensor. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 12661 LNCS:490–502. https://doi.org/10.1007/978-3-030-68763-2_37

Holste G, Wang S, Jiang Z, et al (2022) Long-tailed classification of thorax diseases on chest x-ray: a new benchmark study. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 13567 LNCS:22–32. https://doi.org/10.1007/978-3-031-17027-0_3

Hsu JL, Hung PC, Lin HY et al (2015) Applying under-sampling techniques and cost-sensitive learning methods on risk assessment of breast cancer. J Med Syst. https://doi.org/10.1007/S10916-015-0210-X

Hu K, Huang Y, Huang W et al (2021) Deep supervised learning using self-adaptive auxiliary loss for Covid-19 diagnosis from imbalanced CT images. Neurocomputing 458:232–245. https://doi.org/10.1016/J.NEUCOM.2021.06.012

Huang C, Li Y, Loy CC et al (2020) Deep imbalanced learning for face recognition and attribute prediction. IEEE Transact Pattern Anal Mach Intell 42:2781–2794. https://doi.org/10.1109/TPAMI.2019.2914680

Iranmehr A, Masnadi-Shirazi H, Vasconcelos N (2019) Cost-sensitive support vector machines. Neurocomputing 343:50–64. https://doi.org/10.1016/j.neucom.2018.11.099

ISIC Challenge (2019) ISIC Challenge datasets. Accessed 22 August 2023, https://challenge.isic-archive.com/data/#2019

Javidi M, Abbaasi S, Atashi SN et al (2021) Covid-19 early detection for imbalanced or low number of data using a regularized cost-sensitive capsnet. Sci Rep. https://doi.org/10.1038/S41598-021-97901-4

Jia Y, Shelhamer E, Donahue J, et al (2014) Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on multimedia. Association for Computing Machinery, 14:675-678, https://doi.org/10.1145/2647868.2654889

Jiang J, Liu X, Zhang K et al (2017) Automatic diagnosis of imbalanced ophthalmic images using a cost-sensitive deep convolutional neural network. BioMed Eng Online 16:1–20. https://doi.org/10.1186/S12938-017-0420-1

Jiang Z, Zhao W (2021) Fusion algorithm for imbalanced EEG data processing in seizure detection. Seizure 91:207–211. https://doi.org/10.1016/J.SEIZURE.2021.06.023

Johnson JM, Khoshgoftaar TM (2019) Survey on deep learning with class imbalance. J Big Data 6:1–54. https://doi.org/10.1186/s40537-019-0192-5

Johnson KW, Soto JT, Glicksberg BS et al (2018) Artificial intelligence in cardiology. J Am Coll Cardiol 71:2668–2679. https://doi.org/10.1016/J.JACC.2018.03.521

Kaur H, Pannu HS, Malhi AK (2019) A systematic review on imbalanced data challenges in machine learning. ACM Comput Surv (CSUR). https://doi.org/10.1145/3343440

KEEL (2018) A software tool to assess evolutionary algorithms for data mining problems. Accessed 22 August 2023, http://www.keel.es/

Keras (2023) Keras documentation. Accessed 22 August 2023, https://keras.io/

Khan SH, Hayat M, Bennamoun M et al (2018) Cost-sensitive learning of deep feature representations from imbalanced data. IEEE Transact Neural Netw Learn Syst 29:3573–3587. https://doi.org/10.1109/TNNLS.2017.2732482

Khan Y, Ostfeld AE, Lochner CM et al (2016) Monitoring of vital signs with flexible and wearable medical devices. Adv Mater 28:4373–4395. https://doi.org/10.1002/ADMA.201504366

Khushi M, Shaukat K, Alam TM et al (2021) A comparative performance analysis of data resampling methods on imbalance medical data. IEEE Access 9:109960–109975. https://doi.org/10.1109/ACCESS.2021.3102399

Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering. Tech Rep

Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw. https://doi.org/10.18637/JSS.V028.I05

Kukar M, Kononenko I (1998) Cost-sensitive learning with neural networks. In: European conference on artificial intelligence

Kumar P, Thakur RS (2021) Liver disorder detection using variable- neighbor weighted fuzzy k nearest neighbor approach. Multimed Tools Appl 80:16515–16535. https://doi.org/10.1007/S11042-019-07978-3

König IR, Fuchs O, Hansen G et al (2017) What is precision medicine? Eur Respir J 50:1700391. https://doi.org/10.1183/13993003.00391-2017

Lankireddy P, Sindhura C, Gorthi S (2022) A new lightweight architecture and a class imbalance aware loss function for multi-label classification of intracranial hemorrhages. Lecture notes in computer science (Including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 13583 LNCS:397–405. https://doi.org/10.1007/978-3-031-21014-3_41

Lee CH, Kim HJ, Kim YT et al (2023) Sleepexpertnet: high-performance and class-balanced deep learning approach inspired from the expert neurologists for sleep stage classification. J Ambient Intell Human Comput 14:8067–8083. https://doi.org/10.1007/S12652-022-04443-2

Leevy JL, Khoshgoftaar TM, Bauder RA et al (2018) A survey on addressing high-class imbalance in big data. J Big Data 5:1–30. https://doi.org/10.1186/S40537-018-0151-6

Li H, Xue FF, Chaitanya K et al (2021) Imbalance-aware self-supervised learning for 3d radiomic representations. Lect Notes Comput Sci (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 12902:36–46. https://doi.org/10.1007/978-3-030-87196-3_4

Li H, Dong X, Shen W et al (2022) Resampling-based cost loss attention network for explainable imbalanced diabetic retinopathy grading. Comput Biol Med. https://doi.org/10.1016/J.COMPBIOMED.2022.105970

Li Y, Qian R, Li K (2022) Inter-patient arrhythmia classification with improved deep residual convolutional neural network. Comput Methods Progr Biomed 214:106582. https://doi.org/10.1016/J.CMPB.2021.106582

LightGBM (2023) LightGBM Documentation. Accessed 22 August 2023, https://lightgbm.readthedocs.io/en/stable/

Lili W, Zhongliang F, Pan T (2016) Four-chamber plane detection in cardiac ultrasound images based on improved imbalanced adaboost algorithm. Proceedings of 2016 IEEE international conference on cloud computing and big data analysis, ICCCBDA 2016 pp 299–303. https://doi.org/10.1109/ICCCBDA.2016.7529574

Lin TY, Goyal P, Girshick R et al (2020) Focal loss for dense object detection. IEEE Transact Pattern Anal Mach Intell 42(2):318–327. https://doi.org/10.1109/TPAMI.2018.2858826

Ling CX, Sheng VS (2008) Cost-sensitive learning and the class imbalance problem. Encycl Mach Learn 2011:231–235

Ling CX, Yang Q, Wang J, et al (2004) Decision trees with minimal costs. In: Proceedings of the twenty-first international conference on machine learning. Association for Computing Machinery, New York, NY, USA, ICML '04, p 69, https://doi.org/10.1145/1015330.1015369

Liu N, Shen J, Xu M et al (2018) Improved cost-sensitive support vector machine classifier for breast cancer diagnosis. Math Probl Eng. https://doi.org/10.1155/2018/3875082

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. https://doi.org/10.1016/J.ARTMED.2019.101723

Liu Y, Li Q, Wang K et al (2021) Automatic multi-label ECG classification with category imbalance and cost-sensitive thresholding. Biosensors 11:453. https://doi.org/10.3390/BIOS11110453

Lu Y, Jiang M, Wei L et al (2021) Automated arrhythmia classification using depthwise separable convolutional neural network with focal loss. Biomed Signal Process Control 69:102843. https://doi.org/10.1016/J.BSPC.2021.102843

López V, Fernández A, García S et al (2013) An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf Sci 250:113–141. https://doi.org/10.1016/J.INS.2013.07.007

MATLAB (2023a) Handle imbalanced data or unequal misclassification costs in classification ensembles. Accessed 22 August 2023, https://www.mathworks.com/help/stats/classification-with-unequal-misclassification-costs.html

MATLAB (2023b) MATLAB. Accessed 22 August 2023, https://www.mathworks.com/products/matlab.html

MATLAB (2023c) Train sequence classification network using data with imbalanced classes. Accessed 22 August 2023, https://www.mathworks.com/help/deeplearning/ug/sequence-classification-using-inverse-frequency-class-weights.html

Mazurowski MA, Habas PA, Zurada JM et al (2008) Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. Neural Netw 21:427–436. https://doi.org/10.1016/J.NEUNET.2007.12.031

Mello MM, Lieou V, Goodman SN (2018) Clinical trial participants' views of the risks and benefits of data sharing. New Engl J Med 378:2202–2211. https://doi.org/10.1056/NEJMsa1713258

Mienye ID, Sun Y (2021) Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. Inf Med Unlocked 25:100690. https://doi.org/10.1016/J.IMU.2021.100690

Mirbabaie M, Stieglitz S, Frick NR (2021) Artificial intelligence in disease diagnostics: a critical review and classification on the current state of research guiding future direction. Health Technol 11:693–731. https://doi.org/10.1007/S12553-021-00555-5

Moody G, Mark R (1980) MIT-BIH arrhythmia database. https://doi.org/10.13026/C2F305

Moons KG, Royston P, Vergouwe Y et al (2009) Prognosis and prognostic research: what, why, and how? BMJ 338:1317–1320. https://doi.org/10.1136/BMJ.B375

Munagala NV, Saravanan V, Almukhtar FH et al (2022) Supervised approach to identify autism spectrum neurological disorder via label distribution learning. Comput Intell Neurosci. https://doi.org/10.1155/2022/4464603

Naceur MB, Kachouri R, Akil M et al (2019) A new online class-weighting approach with deep neural networks for image segmentation of highly unbalanced glioblastoma tumors. Lect Notes Comput Sci 11507:555–567. https://doi.org/10.1007/978-3-030-20518-8_46

Naceur MB, Akil M, Saouli R et al (2020) Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy. Med Image Anal 63:101692. https://doi.org/10.1016/J.MEDIA.2020.101692

Nasalwai N, Punn NS, Sonbhadra SK et al (2021) Addressing the class imbalance problem in medical image segmentation via accelerated tversky loss function. Lect Notes Comput Sci 12714:390–402. https://doi.org/10.1007/978-3-030-75768-7_31

Naseem U, Khushi M, Khan SK et al (2020) Diabetic retinopathy detection using multi-layer neural networks and split attention with focal loss. Lect Notes Comput Sci 12534:26–37. https://doi.org/10.1007/978-3-030-63836-8_3

National Institute of Diabetes and Digestive and Kidney Diseases (1990) Pima indians diabetes database. Accessed 22 August 2023, https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Newaz A, Ahmed N, Haq FS (2021) Diagnosis of liver disease using cost-sensitive support vector machine classifier.In:2021 international conference on computational performance evaluation, ComPE 2021 pp 421–425. https://doi.org/10.1109/COMPE53109.2021.9752075

Nunes C, Silva D, Guerreiro M et al (2013) Class imbalance in the prediction of dementia from neuropsychological data. Lect Notes Comput Sci 8154:138–151. https://doi.org/10.1007/978-3-642-40669-0_13

Oracle (2023) Java programming language. Accessed 22 August 2023, https://docs.oracle.com/javase/8/docs/technotes/guides/language/index.html

Ormeño P, Ramírez F, Valle C et al (2012) Robust asymmetric adaboost. Lect Notes Comput Sci 7441:519–526. https://doi.org/10.1007/978-3-642-33275-3_64

Patel H, Rajput DS, Reddy GT et al (2020) A review on classification of imbalanced data for wireless sensor networks. Int J Distrib Sens Netw 16(4):1550147720916404. https://doi.org/10.1177/1550147720916404

Petersen K, Vakkalanka S, Kuzniarz L (2015) Guidelines for conducting systematic mapping studies in software engineering: an update. Inf Softw Technol 64:1–18. https://doi.org/10.1016/j.infsof.2015.03.007

Pranto B, Mehnaz SM, Momen S, et al (2020) Prediction of diabetes using cost sensitive learning and oversampling techniques on bangladeshi and indian female patients.In:Proceedings of ICITR 2020 - 5th international conference on information technology research: towards the new digital enlightenment https://doi.org/10.1109/ICITR51448.2020.9310892

Prashanth R, Roy SD (2018) Novel and improved stage estimation in parkinson's disease using clinical scales and machine learning. Neurocomputing 305:78–103. https://doi.org/10.1016/J.NEUCOM.2018.04.049

Punn N, Agarwal S (2021) Automated diagnosis of Covid-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks. Appl Intell 51:1–14. https://doi.org/10.1007/s10489-020-01900-3

Python (2023) Welcome to python.org. Accessed 22 August 2023, https://www.python.org/

PyTorch (2023a) PyTorch. Accessed 22 August 2023, https://pytorch.org/

PyTorch (2023b) PyTorch Documentation: torch.nn.CrossEntropyLoss. Accessed 22 August 2023, https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html

Qian S, Ren K, Zhang W et al (2022) Skin lesion classification using CNNS with grouping of multi-scale attention and class-specific loss weighting. Comput Methods Progr Biomed 226:107166. https://doi.org/10.1016/J.CMPB.2022.107166

Qin Z, Zhang C, Wang T et al (2010) Cost sensitive classification in data mining. Lect Notes Comput Sci 6440:1–11. https://doi.org/10.1007/978-3-642-17316-5_1

Qin Z, Wang AT, Zhang C et al (2013) Cost-sensitive classification with k-nearest neighbors. In: Wang M (ed) Knowl Sci Eng Manag. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp 112–131

Quinlan R (1987) Thyroid disease. UCI Mach Learn Repos. https://doi.org/10.24432/C5D010

R (2022a) RDocumentation: LiblineaR function. Accessed 22 August 2023, https://www.rdocumentation.org/packages/LiblineaR

R (2022b) RDocumentation; Rpart function. Accessed 22 August 2023, https://www.rdocumentation.org/packages/rpart/

R (2023) R: the R project for statistical computing. Accessed 22 August 2023, https://www.r-project.org/

Rahman A, Hassan I, Ahad MAR (2021a) Nurse care activity recognition: a cost-sensitive ensemble approach to handle imbalanced class problem in the wild. UbiComp/ISWC 2021 - Adjunct proceedings of the 2021 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2021 ACM international symposium on wearable computers pp 440–445. https://doi.org/10.1145/3460418.3479389

Rahman S, Sarker S, Miraj MAA et al (2021) Deep learning-driven automated detection of Covid-19 from radiography images: a comparative analysis. Cogn Comput. https://doi.org/10.1007/S12559-020-09779-5

Raj S, Mahanand BS, Vinod DS (2021) Diffuse lung disease classification based on texture features and weighted extreme learning machine. Multimed Tools Appl 80:35467–35479. https://doi.org/10.1007/S11042-020-10469-5

Rajkomar A, Dean J, Kohane I (2019) Machine learning in medicine. New England J Med 380:1347–1358. https://doi.org/10.1056/NEJMRA1814259

Ramana B, Venkateswarlu N (2012) ILPD (Indian liver patient dataset). UCI Mach Learn Repos. https://doi.org/10.24432/C5D02C

RapidMiner (2023a) Rapidminer | amplify the impact of your people, expertise. Accessed 22 August 2023, https://rapidminer.com/

RapidMiner (2023b) RapidMiner documentation: cost-sensitive scoring. Accessed 22 August 2023, https://docs.rapidminer.com/latest/studio/operators/scoring/cost_sensitive_scoring.html

RapidMiner (2023c) RapidMiner documentation: MetaCost. Accessed 22 August 2023, https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/ensembles/metacost.html

Ravi V (2022) Attention cost-sensitive deep learning-based approach for skin cancer detection and classification. Cancers 14. https://doi.org/10.3390/CANCERS14235872

Ravi V, Narasimhan H, Pham TD (2022) A cost-sensitive deep learning-based meta-classifier for pediatric pneumonia classification using chest x-rays. Exp Syst 39:e12966. https://doi.org/10.1111/EXSY.12966

Razzaghi T, Roderick O, Safro I, et al (2015) Fast imbalanced classification of healthcare data with missing values.In:2015 18th international conference on information fusion, Fusion 2015 pp 774–781

Razzaghi T, Roderick O, Safro I et al (2016) Multilevel weighted support vector machine for classification on healthcare data with missing values. Plos One 11:e0155119. https://doi.org/10.1371/JOURNAL.PONE.0155119

Rekha G, Tyagi AK, Reddy VK (2019) A wide scale classification of class imbalance problem and its solutions: a systematic literature review. J Comput Sci 15:886–929. https://doi.org/10.3844/JCSSP.2019.886.929

Reychav I, Zhu L, McHaney R et al (2019) Real-time survival prediction in emergency situations with unbalanced cardiac patient data. Health Technol 9:277–287. https://doi.org/10.1007/S12553-019-00307-6

Rezaei M, Yang H, Meinel C (2019) Voxel-gan: adversarial framework for learning imbalanced brain tumor segmentation. Lect Notes Comput Sci 11384:321–333. https://doi.org/10.1007/978-3-030-11726-9_29

Roy S, Tyagi M, Bansal V et al (2022) Svd-Clahe boosting and balanced loss function for Covid-19 detection from an imbalanced chest x-ray dataset. Comput Biol Med 150:106092. https://doi.org/10.1016/J.COMPBIOMED.2022.106092

Sadeghi S, Khalili D, Ramezankhani A et al (2022) Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods. BMC Med Inf Decis Making. https://doi.org/10.1186/S12911-022-01775-Z

Sahin Y, Bulkan S, Duman E (2013) A cost-sensitive decision tree approach for fraud detection. Exp Syst Appl 40:5916–5923. https://doi.org/10.1016/J.ESWA.2013.05.021

Scikit-learn (2023a) Scikit-learn Documentation. Accessed 22 August 2023, https://scikit-learn.org/stable/

Scikit-learn (2023b) Scikit-learn Documentation: class_weight.compute_class_weight. Accessed 22 August 2023, https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html

Shan P, Chen J, Fu C et al (2023) Automatic skin lesion classification using a novel densely connected convolutional network integrated with an attention module. J Ambient Intell Human Comput 14:8943–8956. https://doi.org/10.1007/S12652-022-04400-Z

Shen Q, Yang X, Zou L et al (2022) Multitask residual shrinkage convolutional neural network for sleep apnea detection based on wearable bracelet photoplethysmography. IEEE Intern Things J 9:25207–25222. https://doi.org/10.1109/JIOT.2022.3195777

Shen X, Wang G, Kwan RYC et al (2020) Using dual neural network architecture to detect the risk of dementia with community health data: algorithm development and validation study. JMIR Med Inf. https://doi.org/10.2196/19870

Sheng JQ, Hu PJH, Liu X et al (2021) Predictive analytics for care and management of patients with acute diseases: deep learning-based method to predict crucial complication phenotypes. J Med Intern Res. https://doi.org/10.2196/18372

Sheng VS, Ling CX (2006) Thresholding for making classifiers cost-sensitive. In: Proceedings of the 21st national conference on artificial intelligence vol 1. AAAI Press, AAAI'06, pp 476-481

Shi H, Wang H, Huang Y et al (2019) A hierarchical method based on weighted extreme gradient boosting in ECG heartbeat classification. Comput Methods Progr Biomed 171:1–10. https://doi.org/10.1016/J.CMPB.2019.02.005

Shirokikh B, Shevtsov A, Kurmukov A et al (2020) Universal loss reweighting to balance lesion size inequality in 3D medical image segmentation. Lect Notes Comput Sci 12264:523–532. https://doi.org/10.1007/978-3-030-59719-1_51

Siddiqui MK, Huang X, Morales-Menendez R et al (2020) Machine learning based novel cost-sensitive seizure detection classifier for imbalanced EEG data sets. Int J Interact Design Manuf 14:1491–1509. https://doi.org/10.1007/S12008-020-00715-3

Siers MJ, Islam MZ (2020) Class imbalance and cost-sensitive decision trees: a unified survey based on a core similarity. ACM Trans Knowl Discov Data. https://doi.org/10.1145/3415156

Sterner P, Goretzko D, Pargent F (2021) Everything has its price: Foundations of cost-sensitive learning and its application in psychology. [Preprint] PsyArXiv https://doi.org/10.31234/osf.io/7asgz

Sun Y, Kamel MS, Wong AK et al (2007) Cost-sensitive boosting for classification of imbalanced data. Pattern Recogn 40:3358–3378. https://doi.org/10.1016/J.PATCOG.2007.04.009

Sun Y, Wong AK, Kamel MS (2011) Classification of imbalanced data: a review. https://doiorg/101142/S0218001409007326 23:687–719. https://doi.org/10.1142/S0218001409007326

Sung SF, Hung LC, Hu YH (2021) Developing a stroke alert trigger for clinical decision support at emergency triage using machine learning. Int J Med Inf. https://doi.org/10.1016/J.IJMEDINF.2021.104505

Taghanaki SA, Zheng Y, Kevin Zhou S et al (2019) Combo loss: handling input and output imbalance in multi-organ segmentation. Comput Med Imag Graph 75:24–33. https://doi.org/10.1016/j.compmedimag.2019.04.005

Tang Y, Zhang YQ, Chawla NV (2009) Svms modeling for highly imbalanced classification. IEEE Transact Syst Man Cybern Part B Cybern 39:281–288. https://doi.org/10.1109/TSMCB.2008.2002909

Tarekegn AN, Giacobini M, Michalak K (2021) A review of methods for imbalanced multi-label classification. Pattern Recogn 118:107965. https://doi.org/10.1016/J.PATCOG.2021.107965

TensorFlow (2023a) TensorFlow API documentation: weighted_cross_entropy_with_logits. Accessed 22 August 2023, https://www.tensorflow.org/api_docs/python/tf/nn/weighted_cross_entropy_with_logits

TensorFlow (2023b) TensorFlow Documentation. Accessed 22 August 2023, https://www.tensorflow.org/

Ting KM (2002) An instance-weighting method to induce cost-sensitive trees. IEEE Transact Knowl Data Eng 14:659–665. https://doi.org/10.1109/TKDE.2002.1000348

Trigg L (2023a) Costsensitiveclassifier. Accessed 22 August 2023, https://weka.sourceforge.io/doc.dev/weka/classifiers/meta/CostSensitiveClassifier.html

Trigg L (2023b) Metacost. Accessed 22 August 2023, https://weka.sourceforge.io/doc.stable/weka/classifiers/meta/MetaCost.html

Tschandl P (2018) The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. https://doi.org/10.7910/DVN/DBW86T

Turney PD (2002) Types of cost in inductive concept learning. https://arxiv.org/abs/cs/0212034v1

UCI Machine Learning Repository (1990) Liver disorders. Accessed 22 August 2023, https://doi.org/10.24432/C54G67

Uguroglu S, Carbonell J, Doyle M et al (2012) Cost-sensitive risk stratification in the diagnosis of heart disease. Proc Natl Conf Artif Intell 3:2335–2340. https://doi.org/10.1609/AAAI.V26I2.18980

Vanderschueren T, Verdonck T, Baesens B et al (2022) Predict-then-optimize or predict-and-optimize? an empirical evaluation of cost-sensitive learning strategies. Inf Sci 594:400–415. https://doi.org/10.1016/j.ins.2022.02.021

Wang EK, Zhang X, Pan L (2019) Automatic classification of cad ECG signals with ADAE and bidirectional long short-term network. IEEE Access 7:182873–182880. https://doi.org/10.1109/ACCESS.2019.2936525

Wang H, Cui Z, Chen Y et al (2018) Predicting hospital readmission via cost-sensitive deep learning. IEEE/ACM Transact Comput Biol Bioinf 15:1968–1978. https://doi.org/10.1109/TCBB.2018.2827029

Wang KJ, Makond B, Wang KM (2013) An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data. BMC Med Inf Decis Mak 13:1–14. https://doi.org/10.1186/1472-6947-13-124

Wang M, Jiang H, Shi T et al (2022) PSR-nets: deep neural networks with prior shift regularization for PET/CT based automatic, accurate, and calibrated whole-body lymphoma segmentation. Comput Biol Med. https://doi.org/10.1016/J.COMPBIOMED.2022.106215

Wang S, Yao X (2012) Multiclass imbalance problems: analysis and potential solutions. IEEE Transact Syst Man Cybern Part B Cybern 42:1119–1130. https://doi.org/10.1109/TSMCB.2012.2187280

Wang SH, Cheng H, Phillips P et al (2018) Multiple sclerosis identification based on fractional fourier entropy and a modified jaya algorithm. Entropy. https://doi.org/10.3390/E20040254

Wang W, Chakraborty G, Chakraborty B (2020) Predicting the risk of chronic kidney disease (CKD) using machine learning algorithm. Appl Sci 11:202. https://doi.org/10.3390/APP11010202

Wang Y, Wei Y, Yang H et al (2020) Utilizing imbalanced electronic health records to predict acute kidney injury by ensemble learning and time series model. BMC Med Inf Decis Mak 20:1–13. https://doi.org/10.1186/S12911-020-01245-4

Wang YC, Cheng CH (2021) A multiple combined method for rebalancing medical data with class imbalances. Comput Biol Med 134:104527. https://doi.org/10.1016/J.COMPBIOMED.2021.104527

Wang Z, Zhu Y, Li D et al (2020) Feature rearrangement based deep learning system for predicting heart failure mortality. Comput Methods Progr Biomed 191:105383. https://doi.org/10.1016/J.CMPB.2020.105383

WEKA (2023) The WEKA Workbench. Online appendix for data mining: practical machine learning tools and techniques. Accessed 22 August 2023, https://www.cs.waikato.ac.nz/ml/weka/

Wolberg W, Mangasarian O, Street N et al (1995) Breast cancer wisconsin (diagnostic). UCI Mach Learn Repos. https://doi.org/10.24432/C5DW2B

World Health Organization (2016) Mental health and neurological disorders - Q &A. Accessed 22 August 2023, https://www.who.int/news-room/questions-and-answers/item/mental-health-neurological-disorders

World Health Organization (2021) Cardiovascular diseases (CVDs) - Fact Sheet. Accessed 22 August 2023, https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

World Health Organization (2022) Cancer - Fact Sheet. Accessed 22 August 2023, https://www.who.int/news-room/fact-sheets/detail/cancer

Wu JC, Shen J, Xu M et al (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 13:1608–1618. https://doi.org/10.2991/IJCIS.D.201012.005

Wu Y, Pei C, Ruan C et al (2022) Bayesian networks and chained classifiers based on SVM for traditional Chinese medical prescription generation. World Wide Web 25:1447–1468. https://doi.org/10.1007/S11280-021-00981-5

XGBoost (2022) XGBoost Documentation. Accessed 22 August 2023, https://xgboost.readthedocs.io/en/stable/

Xu X, Wang C, Guo J et al (2020) Mscs-deepln: Evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks. Med Image Anal. https://doi.org/10.1016/J.MEDIA.2020.101772

Yang H, Li X, Cao H et al (2021) Using machine learning methods to predict hepatic encephalopathy in cirrhotic patients with unbalanced data. Comput Methods Progr Biomed. https://doi.org/10.1016/J.CMPB.2021.106420

Yao L, Wong PK, Zhao B (2022) Cost-sensitive broad learning system for imbalanced classification and its medical application. Mathematics 829(10):829. https://doi.org/10.3390/MATH10050829

Zeng R, Lu Y, Long S et al (2021) Cardiotocography signal abnormality classification using time-frequency features and ensemble cost-sensitive SVM classifier. Comput Biol Med 130:104218. https://doi.org/10.1016/J.COMPBIOMED.2021.104218

Zhang D, Shen D (2011) Multicost: multi-stage cost-sensitive classification of Alzheimer's disease. Lect Notes Comput Sci 7009:344–351. https://doi.org/10.1007/978-3-642-24319-6_42

Zhang L, Zhao J, Yang H et al (2018) An improved weighted elm with hierarchical feature representation for imbalanced biomedical datasets. Lect Notes Comput Sci 11061:276–283. https://doi.org/10.1007/978-3-319-99365-2_25

Zhang S (2020) Cost-sensitive KNN classification. Neurocomputing 391:234–242. https://doi.org/10.1016/j.neucom.2018.11.101

Zhao H (2008) Instance weighting versus threshold adjusting for cost-sensitive classification. Knowl Inf Syst 15:321–334. https://doi.org/10.1007/S10115-007-0079-1

Zhao R, Chen X, Chen Z et al (2022) Diagnosing glaucoma on imbalanced data with self-ensemble dual-curriculum learning. Med Image Anal 75:102295. https://doi.org/10.1016/J.MEDIA.2021.102295

Zhao Y, Wong ZSY, Tsui KL (2018) A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. J Healthcare Eng 2018:6275435. https://doi.org/10.1155/2018/6275435

Zhao Y, Ren J, Zhang B et al (2023) An explainable attention-based TCN heartbeats classification model for arrhythmia detection. Biomed Signal Process Control. https://doi.org/10.1016/J.BSPC.2022.104337

Zhenya Q, Zhang Z (2021) A hybrid cost-sensitive ensemble for heart disease prediction. BMC Med Inf Decis Mak 21:1–18. https://doi.org/10.1186/S12911-021-01436-7

Zhou B, Yao Y, Luo J (2014) Cost-sensitive three-way email spam filtering. J Intell Inf Syst 42:19–45. https://doi.org/10.1007/S10844-013-0254-7

Zhou X, Hu Y, Liang W et al (2021) Variational lSTM enhanced anomaly detection for industrial big data. IEEE Transact Ind Inf 17:3469–3477. https://doi.org/10.1109/TII.2020.3022432

Zhou ZH, Liu XY (2006) Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transact Knowl Data Eng 18:63–77. https://doi.org/10.1109/TKDE.2006.17

Zieba M (2014) Service-oriented medical system for supporting decisions with missing and imbalanced data. IEEE J Biomed Health Inf 18:1533–1540. https://doi.org/10.1109/JBHI.2014.2322281

Zieba M, Tomczak JM, Lubicz M et al (2014) Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. Appl Soft Comput 14:99–108. https://doi.org/10.1016/J.ASOC.2013.07.016

Zubair M, Yoon C (2022) Cost-sensitive learning for anomaly detection in imbalanced ECG data using convolutional neural networks. Sensors. https://doi.org/10.3390/S22114075