cancers

MDPI

Review

Bias and Class Imbalance in Oncologic Data—Towards Inclusive and Transferrable AI in Large Scale Oncology Data Sets

Erdal Tasci [1,2], Ying Zhuge [1], Kevin Camphausen [1] and Andra V. Krauze [1,*]

[1] Center for Cancer Research, National Cancer Institute, NIH, Building 10, Bethesda, MD 20892, USA; erdal.tasci@nih.gov (E.T.); zhugey@mail.nih.gov (Y.Z.); camphauk@mail.nih.gov (K.C.)
[2] Department of Computer Engineering, Ege University, Izmir 35100, Turkey
[*] Correspondence: andra.krauze@nih.gov

Simple Summary: Large-scale medical data carries significant areas of underrepresentation and bias at all levels: clinical, biological, and management. Resulting data sets and outcome measures reflect these shortcomings in clinical, imaging, and omics data with class imbalance emerging as the single most significant issue inhibiting meaningful and reproducible conclusions while impacting the transfer of findings between the lab and clinic and limiting improvement in patient outcomes. When employing artificial intelligence methods, class imbalance can produce classifiers whose predicted class probabilities are geared toward the majority class ignoring the significance of minority classes, in turn generating algorithmic bias. The inability to mitigate this can guide an AI system in favor of or against various cohorts or variables. We review sources of bias and class imbalance and relate this to AI methods. We discuss avenues to mitigate these and propose a set of guidelines aimed at limiting and addressing data and algorithmic bias.

Abstract: Recent technological developments have led to an increase in the size and types of data in the medical field derived from multiple platforms such as proteomic, genomic, imaging, and clinical data. Many machine learning models have been developed to support precision/personalized medicine initiatives such as computer-aided detection, diagnosis, prognosis, and treatment planning by using large-scale medical data. Bias and class imbalance represent two of the most pressing challenges for machine learning-based problems, particularly in medical (e.g., oncologic) data sets, due to the limitations in patient numbers, cost, privacy, and security of data sharing, and the complexity of generated data. Depending on the data set and the research question, the methods applied to address class imbalance problems can provide more effective, successful, and meaningful results. This review discusses the essential strategies for addressing and mitigating the class imbalance problems for different medical data types in the oncologic domain.

Keywords: class imbalance; oncology; machine learning; artificial intelligence; clinical data

check for updates

Citation: Tasci, E.; Zhuge, Y.; Camphausen, K.; Krauze, A.V. Bias and Class Imbalance in Oncologic Data—Towards Inclusive and Transferrable AI in Large Scale Oncology Data Sets. *Cancers* **2022**, *14*, 2897. https://doi.org/10.3390/cancers14122897

Academic Editors: Alan Hutson and Song Liu

Received: 27 May 2022
Accepted: 9 June 2022
Published: 12 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Large-scale medical data carries significant areas of underrepresentation and bias at all levels: clinical (demographic, disease characteristics), biological (heterogeneity of disease), management and outcomes [1–4]. Underrepresentation or overrepresentation of one or several categories as related to clinical and disease characteristics, management or outcomes, and resulting imaging studies and biospecimens results in data set bias and class imbalance, or unequal distribution of features in a provided data set. This is often exemplified in oncologic data sets by, for example, overrepresentation of patients with superior performance status on clinical trials or, for example, underrepresentation of certain demographics such as the inclusion of certain racial and ethnic groups or patient populations who experience greater challenges with access to care. These uneven distributions within data sets result in a majority class, or the subset of the data that is overrepresented, and a minority class, or a subset of the data that is underrepresented. Class imbalance in clinical, imaging, and omics data sets is the rule, not the exception [5]. In the context of clinical data, this may be the result of the type of practice setting (e.g., publicly funded systems vs. academic vs. private practice), location (urban vs. rural, affluent vs. limited resources), data acquisition (manual vs. batch collection from electronic medical records), data curation (handling missing and inaccurate data). This heterogeneity in "big data" is reflected in publicly available data sets in all types of medical settings and clinical contexts [6,7]. The number and importance of publications about the class imbalance problem have increased over time (see Figure 1) [8]. The variability of data at all levels and its presence in large-scale data sets of all subtypes is emerging as the single most significant issue inhibiting the harnessing of data for meaningful and reproducible conclusions. It impacts the transfer of findings between the lab and the clinic and limits improvement in patient outcomes [9]. However, obtaining balanced data in medical data sets of all subtypes is challenging secondary to limitations in patient numbers, cost, privacy and security of data sharing, and the complexity of generated data, resulting in significant efforts to achieve representative data sets [10,11]. For example, in medical image data sets, the number of one class samples generally exceeds the number of other class samples, causing class imbalance problems which are increasingly a hot topic in machine learning and data mining [10]. While the class with fewer examples represents the minority class, another class represents the majority class. Machine learning models/ artificial intelligence (AI) systems considering all data, therefore, produce biased classifiers whose predicted class probabilities are geared toward the majority class ignoring the significance of minority classes [12,13]. The algorithmic bias here is defined as the prejudice embedded in the decision resulting from an AI system in favor or against a person, group, or thing that is taken into account to be unfair [14]. In other words, algorithmic bias is the difference between the predicted value and the actual value of our model. If the bias is high, the model cannot capture the pattern (e.g., essential features) of data, it misses the relations between the features and targets, and the learning model cannot generate appropriate predictions on testing data. This situation is called underfitting. The decision boundaries created by traditional machine learning models tend to be biased toward the majority class; therefore, the minority classes are most likely misclassified for class imbalanced data sets.

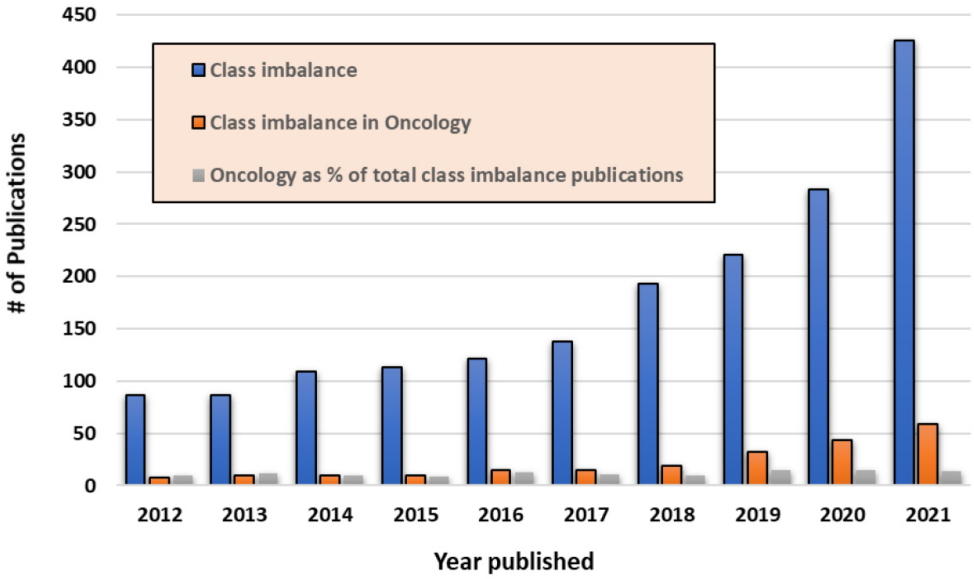

Figure 1. The number of publications on class imbalance listed in Pubmed by year for years 2012 to 2021 (blue = total publications per year on class imbalance, orange = total publications per year on class imbalance in oncology, grey = the number of publications on class imbalance pertaining to oncology as a percent of the total number of publications that year) [8].

Furthermore, the machine learning models treat minority classes as noise; therefore, noise may be incorrectly classified [15]. The degree of imbalance is defined as the ratio of the sample size of the minority class to that of the majority class [11]. The degree of imbalance in medical image data sets generally can be as low as 1:5, 1:10, 1:100, or 1:1000. We will review sources of bias in clinical data sets and their current impact on large scale data use, the effect of data bias and class imbalance on AI methods, and avenues to mitigate these, generating a set of guidelines aimed at limiting and addressing bias and class imbalance in large-scale medical data. The organization of this paper is as follows: Section 1: we discuss sources of data-based class imbalance problems in machine learning as relevant to medical applications (i.e., oncology). Additionally, we provide explanations for the origins of bias and class imbalance in large-scale medical data; Section 2: we discuss methods employed in addressing this problem; Section 3: we present guidelines to mitigate bias and address class imbalance in oncologic data sets toward transferrable AI; Section 4 concludes this paper with future directions.

2. Sources of Bias and Class Imbalance in Large Scale Medical Data

2.1. Gender and Race Imbalance

There is increasing awareness of bias and class imbalance sources in medical data sets. Notably, gender [16,17] and race bias [18,19] with class imbalances are increasingly highlighted in the literature. Both factors have been shown to affect outcomes. They have also been shown to affect data collected at all levels, clinical [6], imaging [13], and biospecimen [20], and in turn, analysis of downstream genomic and proteomic data sets (Figure 2) [8]. Oncologic disciplines where this has been highlighted include large patient volume oncologic sites such as lung cancer [20], breast cancer [18], and head and neck cancer [21] but also lower volume sites such as central nervous system malignancies [22,23].

Nonetheless, in complex data sharing performed with limited resources, a significant proportion of oncologic literature employs large-scale data or uses publicly available data sets for validation, using a biased data set. The use of "easy data" is highly prevalent and results in AI often not transferrable or scalable. Many AI approaches do not employ clinical data or only minimal clinical data, preferring to rely on only imaging or only omic data analyzed with semi-supervised or unsupervised methods. Validation with larger data sets with fewer missing data is carried out in an effort to limit bias. This approach can result in conclusions that have unclear or limited clinical applicability and, due to class imbalance and overfitting, are not replicated in other data sets. Classification of data can improve the outcome of AI approaches but is impacted by the employed data with inter and intra-observer variability.

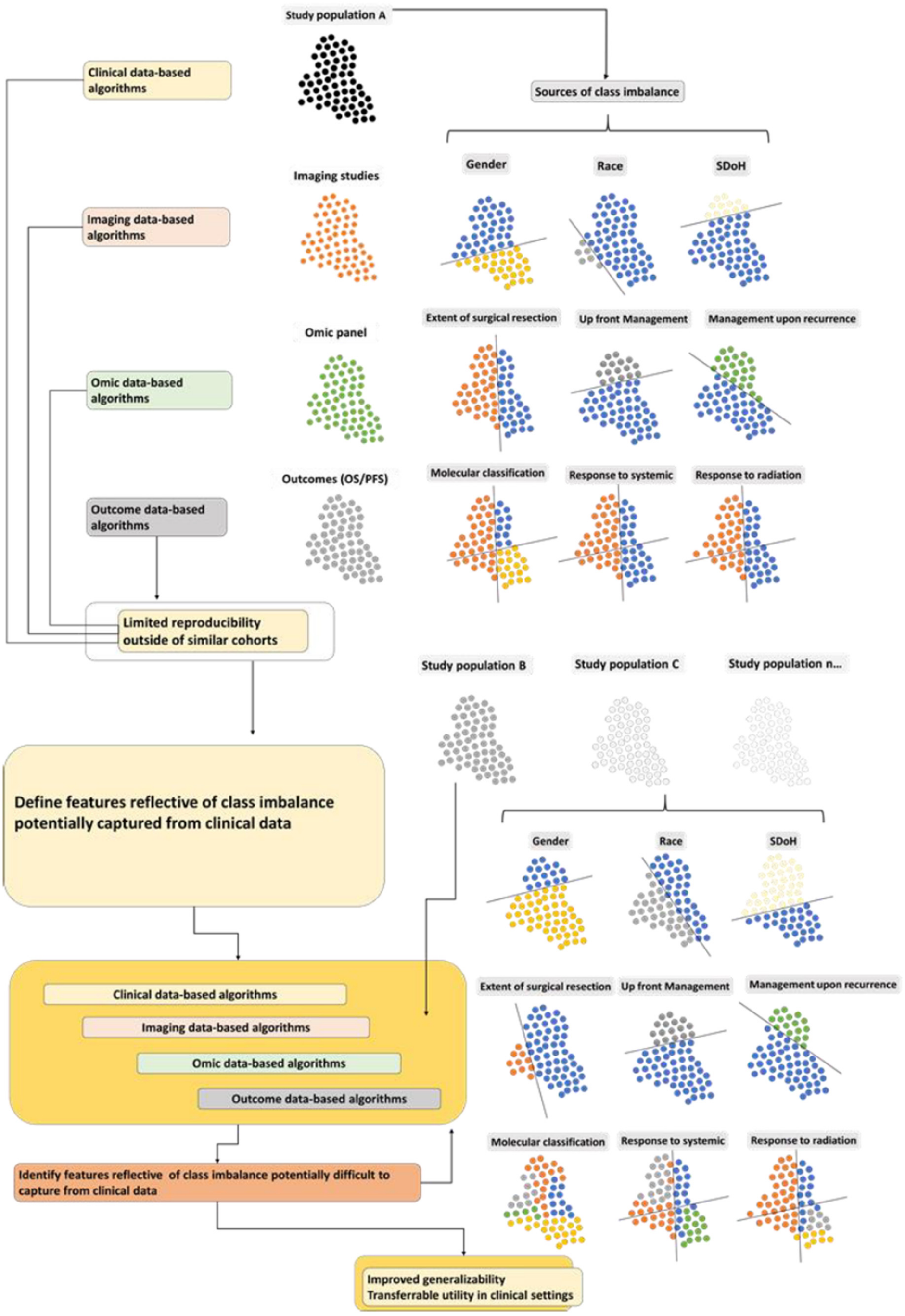
Study population A
Clinical data-based algorithms
Sources of class imbalance
Gender
Race
SDoH
Imaging studies
Imaging data-based algorithms
Extent of surgical resection
Up front Management
Management upon recurrence
Omic panel
Omic data-based algorithms
Outcomes (OS/PFS)
Molecular classification
Response to systemic
Response to radiation
Outcome data-based algorithms
Limited reproducibility outside of similar cohorts
Study population B
Study population C
Study population n...
Define features reflective of class imbalance potentially captured from clinical data
Clinical data-based algorithms
Imaging data-based algorithms
Omic data-based algorithms
Outcome data-based algorithms
Identify features reflective of class imbalance potentially difficult to capture from clinical data
Improved generalizability
Transferrable utility in clinical settings

Figure 2. Clinical, imaging, omic and outcome based algorithms (**left panel**) are generated based on data sets that harbor underrepresentation or overrepresentation of one or several categories as related to demographics (gender, race, social determinants of health–(**right top panel**), management (extent of surgical resection, type of upfront management, management upon recurrence (**right middle panel**) and disease characteristics (molecular classification, response to systemic management and radiation therapy (**right lower panel**). Class imbalance affects ancillary data sets such as imaging, omics based on biospecimens and outcomes (**middle panel**). The resulting algorithms result in limited reproducibility in other cohorts (Study population, B, C, and a theoretical infinite other population that does not share identical class imbalance). As both the sources and extent of over and underrepresentation of certain classes and class imbalance are altered in other populations/data sets (**lower panel**), defining features reflective of class imbalance and mitigating these via compensatory methods in all data subtypes, can address the lack of reproducibility and help identify additional features that can be used to further optimize algorithms allowing for transferrable results (**lower panel**).

2.2. Limited Capture and Inclusion of Social Determinants of Health (SDOH) and Disability

Additional significant sources of data-related bias involve class imbalance due to the lack of capture and inclusion of social determinants of health (SDOH) [24] and patient disability in AI methods. Social determinants of health consist of the geographical and socioeconomic conditions where people are born, live, and age, and the drivers of these conditions with cancer outcome disparities reflecting these aspects [24]. Disability affects the aggressiveness and accessibility of oncologic management but can be challenging to collect consistently in data sets and is intersectional with other factors noted above, such as gender and SDOH. Both SDOH and disability have been particularly challenging to capture robustly in data sets and, as a result, are often not included in AI analysis [24,25]. Class imbalance carries significant implications concerning these domains, and studies have looked at computational approaches, including natural language processing in unstructured data to expand data sets to capture disability as a feature [26].

2.3. Variability in Oncologic Management

Patient management in oncology can vary based on several clinical and disease-related factors, including features discussed in Sections 1 and 2 above and the evolution of staging and treatment paradigms in any one condition over time. Notably, demographics and management can be drastically different between trial populations and the real world [27,28], resulting in heterogeneous data mirrored in data sets employed for analysis, training, testing, and validation (Figure 2). In real-world settings, oncologic management can be highly variable based on geographical location, access, and resources, which can significantly impact training AI algorithms and result in non-transferable products.

2.4. Variability in Oncologic Outcomes Capture

Outcome measures of most studies depend on capturing progression and survival parameters with patient- and physician-reported outcomes increasingly being implemented. When patients are treated in clinical trials, these parameters are more rigorously captured than patient outcome capture in real-world settings. In real-world settings, patient survival (date of diagnosis to the date of death) is a robust parameter (albeit with the limitation of the cause of death being inconsistently captured). However, the date of disease progression is inconsistently available, suffering from difficulties in determining progression (clinical, radiographic, biochemical) and resource limitations to data annotation. Protocols are also attempting to address this aspect with adaptive trial designs that are increasingly complex. Provided the growing cost of developing and testing oncologic management avenues, the ability to identify subpopulations of patients in whom treatment is effective vs. those where early intervention or alteration in management may be beneficial is driven by robust data that originates from highly curated trial populations. Computerized patient-reported outcome measures and patient-generated data are on the rise [29] in clinical settings, although such large-scale data are likely to intersect with other sources of data imbalance. As a result, large-scale medical data are less likely to originate from study patients (which is a source of class imbalance in and of itself) and more likely to grow using real-world data (also a potential source of class imbalances).

2.5. Knowledge Translation Bias

While many a conversation involves direct data bias, as described above, it should be noted that data sets, their analysis, and prediction results, as can be accessed in the public domain, lead back to published results. However, most published studies report positive findings and far less report on unsuccessful AI as exemplified in the context of, for example, skin cancer classification [30,31], sepsis prediction [32], the use of race correction in clinical algorithms [33], and notably the COVID-19 pandemic [34]. Publication bias is well described reflecting both the excitement and the fundamental clinical gaps in aspects of diagnosis, response, or management that need to be overcome with the perceived advantage of artificial intelligence approaches. Negative studies have the potential to be highly instructive but lack publication exposure. The reason for unsatisfactory results may relate to a lack of capture and adaptation to class imbalance sources, limiting the exposure of such studies to other investigators, and the development of meaningful solutions to flawed data.

3. Overview of the Existing Methods for the Class Imbalance Problem

Many studies have been conducted in the literature to address the problem of bias and class imbalance and increase the learning performance of the model utilized. Although there is no universal method to deal with the class imbalance problem, researchers have suggested various methods and approaches for this, according to the size of the data set they use, data distribution, imbalance ratio, and performance criteria of the model. According to the methods used, the existing approaches can be mainly divided into a data level, algorithm level, and hybrid. Additionally, these levels can be grouped into various techniques (i.e., re-sampling, recognition-based approach, cost-sensitive learning, and ensemble learning) [11]. An overview of the existing methods for the class imbalance problem is illustrated in Table 1 and Figure 3. This review summarizes the current technical tools for handling class imbalance problems, particularly medical image data sets. We mainly focus on image-centric issues and the strategies for addressing these in this study. The advantages and drawbacks of the methods employed in the literature are described in the following sections.

Table 1. Pertinent literature addressing approaches to class imbalance.

Study	Technique	Setting
	Data-level methods	
Bose et al. [35]	Convolutional neural networks (CNN), long short-term memory (LSTM), over-sampling	Radiation Oncology, incident reporting
Brown et al. [36]	Under-sampling	Radiation Oncology, prostate cancer
Liu et al. [37]	Synthetic Minority Oversampling Technique (SMOTE)	Glioblastoma imaging
Suarez-Garcia et al. [38]	Under-sampling	Glioma imaging
Li et al. [39]	Under-sampling, K-Means++ and learning vector quantization (LVQ)	Liver cancer
Isensee et al. [40]	Deep learning, nnU-net	Brain tumor segmentation
	Algorithm level methods	
Goyal et al. [41]	Recognition-based, one class classification	real-world data sets across different domains: tabular data, images (CIFAR and ImageNet), audio, and time-series
Gao et al. [42]	Recognition-based, one class classification	Medical imaging
Welch et al. [43]	Recognition-based	Head and Neck Radiation therapy
Leevy et al. [44]	Cost-sensitive	General
Nguyen et al. [45]	Cost-sensitive, comparison of techniques	SMOTE and Deep Belief Network (DBN) against the two cost-sensitive learning methods (focal loss and weighted loss) in churn prediction problem
Milletari et al. [46]	Cost-sensitive	Medical image segmentation
Lin et al. [47]	Cost-sensitive, Focal loss	Medical imaging
Jaeger et al. [48]	Cost-sensitive, Focal loss	Medical imaging object detection

Table 1. *Cont.*

Study	Technique	Setting
Xiong et al. [49]	Cost-Sensitive Naive Bayes Stacking Ensemble	Various malignancy data sets and data types
Shon et al. [50]	Cost-sensitive	Kidney cancer data (TCGA)
Dong et al. [51]	Ensemble-Learning	General
Sagi et al. [52]	Ensemble-Learning	General
Tang et al. [53]	Ensemble-Learning, bagging-based	Transcriptome and functional proteomics data breast cancer
Le et al. [54]	Ensemble-Learning, ResNet50 CNN	Skin cancer
Wang et al. [55]	Ensemble-Learning, multi-layer perceptron (MLP)-based	Gastric cancer
	Hybrid methods	
Khushi et al. [15]	Comparative Performance Analysis of Data Re-sampling Methods	Various malignancy data sets and data types
Chen et al. [56]	Combination of methods	General
Zhao et al. [57]	Random Under-Sampling Boost (RUSBoost)	Colorectal cancer, microarray data
Urdal et al. [58]	Random Under-Sampling Boost (RUSBoost)	Urothelial carcinoma, histopathology data
	Other methods and reviews	
Mirza et al. [59]	Integrative analysis of biomedical big data	
Guyon et al. [60]	variable and feature selection overview	
Hilario et al. [61]	Class imbalance in proteomic biomarker studies overview	
Tibshirani et al. [62]	LASSO (Least Absolute Shrinkage and Selection Operator) overview	
Yan et al. [63]	Graph- and kernel-based—omics data integration algorithms	
Fawcett et al. [64]	ROC analysis overview	

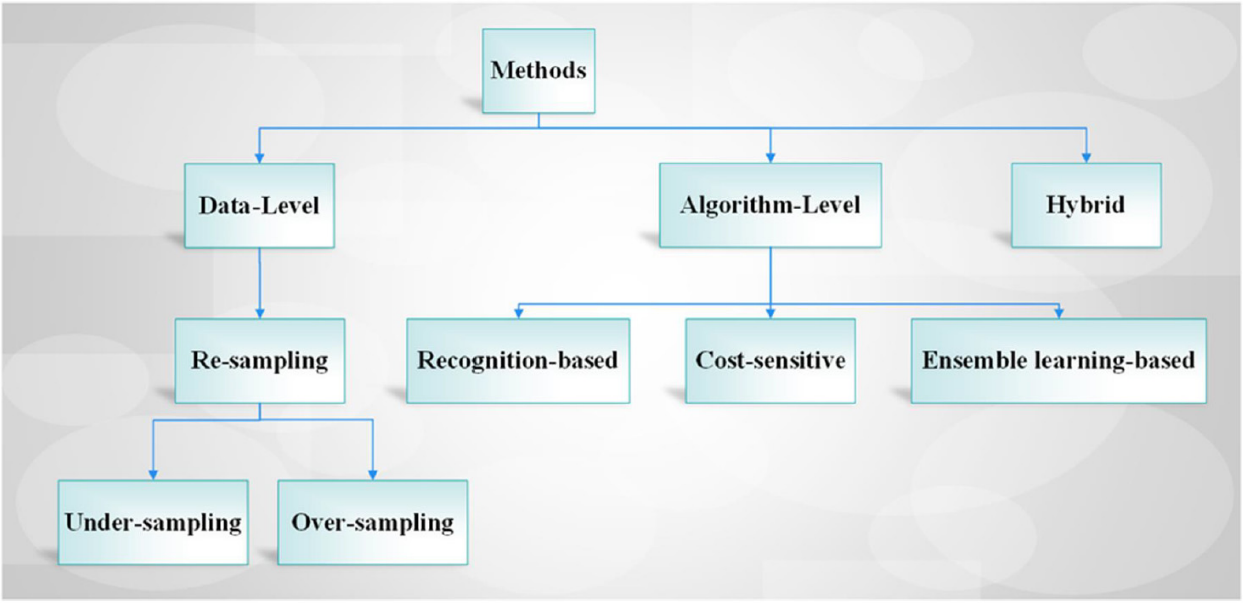

Figure 3. An overview of the existing methods for the class imbalance problem.

3.1. Data-Level Methods

Data level methods can be re-sampling methods such as over-sampling and under-sampling (Figure 3). They aim to balance data sets by utilizing re-sampling methods such as increasing the sample size of the minority classes (over-sampling) or reducing the sample size of the majority class (under-sampling) [15,65]. In some cases, over-sampling methods may cause overfitting due to duplicating a minority of samples for the related data set; under-sampling also may have some disadvantages, such as discarding useful information for the data distribution [15]. Under-sampling can remove valuable and significant patterns; therefore, it can cause the loss of useful information [10]. Both over-sampling and under-sampling methods can be applied separately or as a hybrid (e.g., SMOTE-ENN or SMOTE-Tomek [15]) to any learning model for class imbalance problems to handle these limitations and provide good generalization ability of the model. Both over-sampling and under-sampling methods can be applied separately or as a hybrid to any learning model for class imbalance problems. Data-level methods are a type of data preprocessing stage that redistributes training instances to establish well-balanced data sets. Data-level methods' advantages are independent of underlying the learning model/classifier and can be easily performed [10]. However, under-sampling can remove valuable and significant patterns; therefore, it can cause the loss of useful information [10].

On the other hand, over-sampling results in a time-consuming process due to extra computational cost to the learning model used. To automate the incident triage and severity prioritization process for saving time and cost as well as overcoming errors in the radiation oncology incident learning system, Bose et al. applied convolutional neural networks (CNN) and long short-term memory (LSTM) with minority and random over-sampling strategies to handle high-class imbalanced data [35]. They employed incident reports obtained from Virginia Commonwealth University (VCU) and the US Veterans Health Affairs (VHA) radiation oncology centers. Brown et al. developed various data-filtering techniques, including under-sampling methodologies for imbalanced data classification methods, to provide automated radiation therapy (RT) quality assurance (QA) for prostate cancer treatment. Their proposed approach obtains strong performance results with high classification accuracy [36]. Liu et al. used Synthetic Minority Oversampling Technique (SMOTE) to overcome with class imbalance problem for Glioblastoma Multiforme (GBM) brain tumor images obtained from Moffitt Cancer Research Center and TCGA (The Cancer Genome Atlas Program) T1-Contrast data set (NIH's Cancer Imaging Archive Dataset) [37]. Their method achieved 37.83% AUC (Area under the Curve) improvement for brain magnetic resonance imaging (MRI) images. In another study, Suarez-Garcia et al. also developed a basic model for glioma grading based on texture analysis using the under-sampling method to handle class imbalance data [38].

Furthermore, to predict the occurrence of liver cancer with skewed epidemiological data and tackle the class imbalance problem, Li et al. introduced two different under-sampling methods based on K-Means++ and learning vector quantization (LVQ) [39]. Isensee et al. developed a fast and effective deep learning-based segmentation method, namely nnU-net, which automatically configures all segmentation architecture without user interventions for arbitrary new tasks/data sets [40]. Their method employs an oversampling strategy by randomly selecting 66.7% of samples from the training data for a robust handling class imbalance problem. In case the re-sampling methods may not provide performance improvement of the prediction models, other methods such as ensemble learning (e.g., boosting) or various feature selection approaches can be employed.

3.2. Algorithm-Level Methods

Researchers can alter traditional machine learning models at the algorithm level, assigning weight or cost to a classifier to reduce bias towards the majority class. The learning model is not affected by the class distribution in this process. These methods may be recognition-based, cost-sensitive, or ensemble learning-based (Table 1 and Figure 3).

3.2.1. Recognition-Based Methods

One-class learning is an approach that is used in the absence of non-target class instances for recognition-based methods. Outlier detection and novelty detection are examples of the one-class classification. This method provides high performance, particularly on high-dimensional data. Various models, such as support vector machine (SVM) and isolation forest, can be built by one class learning, while some models, such as decision trees and Naïve Bayes, cannot [41]. Gao et al. proposed a deep-learning-based one-class classification approach to handle imbalanced class data for medical images such as breast MRIs and space-occupying kidney lesion (SOKL) data sets [42]. For head and neck cancer radiation therapy outcome predictions, Welch et al. (2020) employed the isolation forest algorithm, an ensemble of isolation trees constructed to deal with the class imbalance and detect data anomalies [43].

3.2.2. Cost-Sensitive Methods

Cost-sensitive methods are particularly critical for medical applications due to the importance of false positive and false negative instances. In this approach, the misclassification cost is adjusted to balance the majority and minority classes (e.g., a false negative prediction may be assigned a higher weight (i.e., cost) as compared to a false positive prediction [44]. This practical solution yields cost-sensitive learning for this problem. There are various cost-based approaches such as weighted cross-entropy, multiclass dice loss function, and focal loss to overcome the class imbalance in data in the literature.

The weighted cross-entropy loss equation [45] is defined as Equation (1). Cross-entropy loss function, called logarithmic loss, or log loss, is based on the Information Theory and related to the level of uncertainty (i.e., entropy) for the outcome of variables. Cross-entropy is used for adjusting model weights during the training stage to minimize this loss value, and this score is calculated by penalizing the probability based on how far the predicted value is from the true expected value. The smaller the loss indicates, the better the model.

$$L_w = -\sum_{i=1}^{m} \left(\alpha y_i \log(\widehat{y}_i) + (1 - y_i)\log(1 - \widehat{y}_i)\right) \tag{1}$$

where α represents the imbalance parameter.

Another class loss function to handle class imbalance in data for the medical image segmentation is the multiclass dice loss function [46]. It is differentiable and can be easily adapted into various deep learning frameworks. The multiclass dice loss function is illustrated in Equation (2).

$$L_{dc} = -\frac{2}{|K|}\sum_{k \in K} \frac{\sum_i u_{i,k} v_{i,k}}{\sum_i u_{i,k} + \sum_i v_{i,k}} \tag{2}$$

where u denotes the softmax output of the network and v denotes a one-hot encoding of the ground truth segmentation map. u and v have the shape I by K with $i \in I$ being the voxels in the training data and $k \in K$ being the classes. $u_{i,k}$ and $v_{i,k}$ represent the softmax probability output and ground truth for class k at voxel i, respectively [40].

In addition to the weighted cross-entropy loss and multiclass dice loss function, Lin et al. introduced the focal loss method to add a modulating factor (1-p_i)$^{\gamma}$ to the cross-entropy loss with tunable focusing parameter $\gamma \geq 0$ [47]. Focal loss is the improved version of the cross-entropy function. It assigns more weights to misclassified examples (e.g., background image) by increasing the importance of correcting. It is defined as Equation (3). Focal loss is used for semantic segmentation and medical object detection [48].

$$FL(p_t) = -(1 - p_t)^{\lambda} \log(p_t) \tag{3}$$

In a study, the embedding cost-sensitive naive Bayes (NB) classifier was used as a meta-learner to handle class imbalance for cancer gene expression data [49]. In another study, Shon et al. proposed a cost-sensitive hybrid deep learning approach for kidney cancer classification [50].

3.2.3. Ensemble Learning-Based Methods

Ensemble learning combines multiple weak learners to acquire a robust model with better performance by using various mechanisms such as voting and boosting [51,52]. A decision is typically generated in supervised pattern recognition and machine learning tasks [52]. Ensemble learning models have a better performance than individual learning models, and they provide more resistance to noise/outlier. However, the training time for these learning models is time-consuming, and they can cause over-fitting in some cases [10].

Tang et al. constructed multi-omics data with the transcriptome and functional proteomics data and employed a bagging-based ensemble learning method on breast cancer to avoid over-fitting problems [53]. Le et al. utilized an ensemble of pre-trained ResNet50 CNN models for the classification of skin cancer to support dermatologists in skin cancer diagnosis [54]. Their model is based on a class-weighted learning approach for different classes in the loss function. Meanwhile, Wang et al. introduced a multi-layer perceptron (MLP)-based ensemble learning architecture for the classification of advanced gastric cancer (AGC) [55]. To reduce the class imbalance and keep the network compatible, they optimized the MLP with cross-entropy loss.

3.3. Hybrid Methods

Hybrid methods combine data-level methods and algorithm-level methods [15]. Data-level methods are utilized to process data externally and set the distribution of classes for instances. Afterward, the learning process is carried out internally with algorithm-level methods. With hybrid methods, the learning model will not be biased too much towards the majority class during the classification stage [56]. Detailed descriptions of these methods can be found in [15].

Random Under-Sampling Boost (RUSBoost) is one example of a hybrid method. To increase the predictive performance of colorectal cancer with the imbalanced data, Zhao et al. employed the RUSBoost algorithm for the colorectal cancer diagnosis [57]. A RUSBoost classifier was also used by Urdal et al. to solve and alleviate the class imbalance problems for the prognosis prediction of superficial urothelial carcinomas [58].

3.4. Other Methods

For omics and proteomic data in bioinformatics, machine learning-based models are generally faced with the rarity in the target class or class imbalance problem [59]. For example, a predictive model trained to predict the location of enhancers (i.e., positive samples) in the genome suffers from the class imbalance problem due to a large proportion of negative samples (i.e., non-enhancers) [59]. The prediction of contact-map or post-translation modifications (PTM) sites in a protein sequence, DNA methylation status/site prediction, identification of antimicrobial peptides (AMP) functional types, and the mutation incidence are one of the examples of class imbalance problems [59].

Different feature selection approaches and/or ensemble learning methods are employed to resolve this data imbalance problem. Feature selection approaches are considered dimensionality reduction techniques. Due to the high dimensionality of proteomic data, analyzing, storing, training, and classifying these data may be considered an NP-Hard problem [53,60]. Feature selection methods to find optimal feature subsets reduce the computational and storing cost and eliminate redundant and irrelevant information, facilitating data visualization. Several studies in oncology have adopted feature selection for cancer detection or diagnosis. Feature selection methods generally can be grouped into three categories: Filter, wrapper, and embedded methods (see Figure 4). Filter methods are independent of the chosen learning model, selecting variables by utilizing various statistical measures such as correlation. Wrapper methods assess the features according to the performance of the estimated model and selection criterion [61]. Embedded feature selection methods such as least absolute shrinkage and selection operator (LASSO) regression [62] perform the feature selection process as an integral part of the learning process, often specific to a provided predictor. Lao et al. proposed a deep learning-based radiomics model for survival prediction of Glioblastoma Multiforme (GBM) [66], wherein deep and handcrafted features were extracted from multi-modality MR images employing the LASSO regression model for the feature selection. Their deep imaging feature-based biomarker provides a potential for the preoperative care of GBM patients [66]. Tang et al. employed multi-omics data (transcriptome profiles and functional proteomics data) for breast cancer diagnosis [53], introducing a novel framework with hybrid feature selection and bagging-based ensemble learning methods to resolve the class imbalance problem. Wu et al. demonstrated that dosiomics improves predicting locoregional recurrence for head and neck cancer cases and should be proposed correlatively. They constructed the RadModel and RadDosModel for this purpose [67], extracted shape-based, statistical, texture-based radiomic features using Pyradiomics tool, and subsequently selected relevant and essential features which were then compressed using principal component analysis technique (PCA). Ensemble learning technique (EasyEnsemble) was applied to handle class imbalance problems.

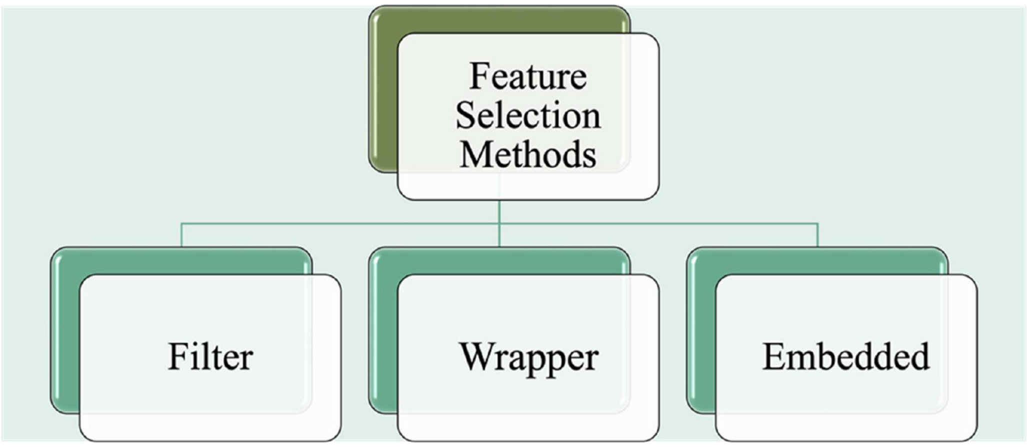

Figure 4. The overview of the feature selection methods.

Apart from these feature selection and ensemble learning methods, various class imbalance-specific evaluation measures are also being proposed. Conventional evaluation measures such as accuracy may not yield effective results in solving class imbalance problems as models may prove insufficiently trustworthy. For example, the accuracy rate may result in high classification performance in class imbalance problems in the event of the presence of a few rare samples, but it may not correctly classify these correctly. Thus, it may be necessary to assign more weight to false positive samples or false negative samples to improve the performance of the predictor instead of relying only on the accuracy rate. Class imbalance-specific evaluation measures such as the area under the precision-recall curve (auPRC), Matthews correlation coefficient (MCC), F-score, and the geometric mean (Gmean) of sensitivity and specificity [59,63] however may not yield adequate results. The equations for accuracy (*ACC*), precision (*PRE*), recall (*REC*), and F-score measures [64] are illustrated in Equation (4).

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad PRE = \frac{TP}{TP + FP} \tag{4}$$

$$REC = \frac{TP}{TP + FN} \qquad F - measure = \frac{2}{1/PRE + 1/REC}$$

where *TP*, *TN*, *FP*, and *FN* represent the true positives, true negatives, false positives, and false negatives, respectively.

4. Guidelines to Mitigate Bias and Address Class Imbalance in Oncologic Data Sets towards Transferrable AI

4.1. Clinical Data

(1) Curate data with a view towards aggregation with parallel data sources that include the possibility to cross-check data collection against other data parameters, e.g., state or government data sources that can act as data verifiers of SDOH, disability, geographical location, and access to care.

(2) Collect binary data when collecting more detailed data is not possible or impractical to obtain.

(3) Collect data electronically with limited manual input and, when possible, automatically limit dependence on clinical provider input to allow for consistent long-term data collection as alternatives augment bias when some centers/settings collect data and others less so or not at all.

(4) Implement data curation/clean data and data “health checks” for known sources of bias and class imbalance.

(5) Implement capture and analysis of SDOH, disability, sex differences, ethnicity, and race data whenever possible and subject analysis to bias check, which can then be accounted for in future research and interpretation of data and resulting algorithms.

4.2. Other Data

Data-level, algorithm-level, and hybrid-level methods can mitigate bias and address class imbalance in oncologic data sets for imaging and omic, proteomic, and genomic data types. Researchers have at their disposal various feature selection/reduction techniques or ensemble learning-based methods that can be utilized for various types of bioinformatics data. In addition to this, diverse performance metrics such as the area under the precision–recall curve (auPRC), Matthews correlation coefficient (MCC), F-score, and the geometric mean (Gmean) of sensitivity and specificity can yield more effective results for class imbalance problems and should be considered when class imbalance is apparent. The most optimal method to be employed for a specific problem will ultimately depend on the data set characteristics (i.e., imbalance ratio) and the problem domain (e.g., the degree of importance, the main objective). Several approaches need to be examined to arrive at the best approach, particularly when a novel hypothesis or novel data set is involved, and the data supporting the most optimal approach may be lacking. We advise experimenting with different combinations, including traditional approaches, before arriving at the most optimal scheme in order to obtain more robust, efficient transferrable results for the related model and data set. However, if the results do not improve, data augmentation/reduction techniques can also be applied.

5. Conclusions

Acknowledging the complexity of the technology required to advance AI in medicine, clinicians represent a set of stakeholders most likely to appreciate potential sources of information bias in data sets. However, technical resources are most likely to possess the capability to mitigate algorithmic bias, while both disciplines are invariably involved in recognizing and addressing pervasive knowledge translation bias. This review examined and summarized diverse methods to handle class imbalance problems for AI in medical settings, particularly oncologic data. Some methods rely on accounting for data distribution, while others relate to modifying the learning model. Because there is no universally accepted solution for all problems, every approach has advantages and disadvantages related to the data set, research question, and learning model used. As a future direction of this work, we aim to improve computer-aided diagnosis (CADx) results and the management of malignancy by employing different and efficient approaches for handling the class imbalance problems in AI. CADx provides unique opportunities to identify biological triggers for tumor response and resistance by leveraging large scale data that may connect to existing scientific knowledge in fascinating and practice changing ways while also posing great challenges to understanding a wealth of growing data while filtering meaningful signals from the noise while conserving and growing crucial pattern recognition to truly positively alter outcomes.

Author Contributions: Conceptualization, E.T. and A.V.K.; Methodology, E.T. and A.V.K.; Investigation, E.T., Y.Z., A.V.K.; Writing—Original draft preparation, E.T., A.V.K.; Writing—Review and editing, E.T., A.V.K., Y.Z., K.C.; Visualization, E.T. and A.V.K.; Supervision, A.V.K. and K.C.; Project administration, A.V.K.; Funding acquisition, A.V.K., K.C. All authors have read and agreed to the published version of the manuscript.

Funding: Funding provided in part by NCI NIH intramural program (ZID BC 010990).

Conflicts of Interest: The authors declare that they have no conflict of interest.

Abbreviations

ACC	Accuracy
AGC	Advanced Gastric Cancer
AI	Artificial Intelligence
AMP	Antimicrobial Peptides
AUPRC	Area Under Precision–Recall Curve
CADx	Computer-Aided Diagnosis
CNN	Convolutional Neural Networks
GBM	Glioblastoma Multiforme
GMEAN	Geometric Mean
LASSO	Least Absolute Shrinkage and Selection Operator
LSTM	Long Short-Term Memory
LVQ	Learning Vector Quantization
MCC	Matthews Correlation Coefficient
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
NB	Naive Bayes
QA	Quality Assurance
PCA	Principal Component Analysis
PRE	Precision
PTM	Post-Translation Modification
REC	Recall
RT	Radiation Therapy
RUSBoost	Random Under-Sampling Boost
SDOH	Social Determinants of Health
SMOTE	Synthetic Minority Oversampling Technique
SOKL	Space-Occpying Kidney Lesion
SVM	Support Vector Machine
VCU	Virginia Commonwealth University
VHA	Veterans Health Affairs

References

1. Belenguer, L. AI bias: Exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI Ethic* **2022**, 1–17. [CrossRef] [PubMed]
2. Cho, M.K. Rising to the challenge of bias in health care AI. *Nat. Med.* **2021**, *27*, 2079–2081. [CrossRef] [PubMed]
3. Daneshjou, R.; Smith, M.P.; Sun, M.D.; Rotemberg, V.; Zou, J. Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms. *JAMA Dermatol.* **2021**, *157*, 1362. [CrossRef] [PubMed]
4. Noor, P. Can we trust AI not to further embed racial bias and prejudice? *BMJ* **2020**, *368*, m363. [CrossRef] [PubMed]
5. Megahed, F.M.; Chen, Y.-J.; Megahed, A.; Ong, Y.; Altman, N.; Krzywinski, M. The class imbalance problem. *Nat. Methods* **2021**, *18*, 1270–1272. [CrossRef]
6. Flynn, E.; Chang, A.; Altman, R.B. Large-scale labeling and assessment of sex bias in publicly available expression data. *BMC Bioinform.* **2021**, *22*, 168. [CrossRef]

7. Williamson, C.W.; Nelson, T.; Thompson, C.A.; Vitzthum, L.K.; Zakeri, K.; Riviere, P.; Bryant, A.K.; Sharabi, A.B.; Zou, J.; Mell, L.K. Bias Reduction through Analysis of Competing Events (BRACE) Correction to Address Cancer Treatment Selection Bias in Observational Data. *Clin. Cancer Res.* **2022**, *28*, 1832–1840. [CrossRef]
8. National Library of Medicine. Available online: https://pubmed.ncbi.nlm.nih.gov (accessed on 1 December 2021).
9. Landers, R.N.; Behrend, T.S. Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *Am. Psychol.* **2022**. [CrossRef]
10. Abd Elrahman, S.M.; Abraham, A. A Review of Class Imbalance Problem. *J. Netw. Innov. Comput.* **2013**, *1*, 332–340.
11. Nguyen, G.H.; Bouzerdoum, A.; Phung, S.L. Learning Pattern Classification Tasks with Imbalanced Data Sets. In *Pattern Recognition*; Yin, P.-Y., Ed.; IntechOpen: Vienna, Austria, 2009. [CrossRef]
12. Rajaraman, S.; Ganesan, P.; Antani, S. Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. *PLoS ONE* **2022**, *17*, e0262838. [CrossRef]
13. Larrazabal, A.J.; Nieto, N.; Peterson, V.; Milone, D.H.; Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Natl. Acad. Sci.* **2020**, *117*, 12592–12594. [CrossRef] [PubMed]
14. Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdl, W.; Vidal, M.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; et al. Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Min. Knowl. Discov.* **2020**, *10*, e1356. [CrossRef]
15. Khushi, M.; Shaukat, K.; Alam, T.M.; Hameed, I.A.; Uddin, S.; Luo, S.; Yang, X.; Reyes, M.C. A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. *IEEE Access* **2021**, *9*, 109960–109975. [CrossRef]
16. Marcu, L.G. Gender and Sex-Related Differences in Normal Tissue Effects Induced by Platinum Compounds. *Pharmaceuticals* **2022**, *15*, 255. [CrossRef] [PubMed]
17. Özdemir, B.; Oertelt-Prigione, S.; Adjei, A.; Borchmann, S.; Haanen, J.; Letsch, A.; Mir, O.; Quaas, A.; Verhoeven, R.; Wagner, A. Investigation of sex and gender differences in oncology gains momentum: ESMO announces the launch of a Gender Medicine Task Force. *Ann. Oncol.* **2022**, *33*, 126–128. [CrossRef] [PubMed]
18. Marcelino, A.C.; Gozzi, B.; Cardoso-Filho, C.; Machado, H.; Zeferino, L.C.; Vale, D.B. Race disparities in mortality by breast cancer from 2000 to 2017 in São Paulo, Brazil: A population-based retrospective study. *BMC Cancer* **2021**, *21*, 998. [CrossRef]
19. Morshed, R.A.; Reihl, S.J.; Molinaro, A.M.; Kakaizada, S.; Young, J.S.; Schulte, J.D.; Butowski, N.; Taylor, J.; Bush, N.A.; Aghi, M.K.; et al. The influence of race and socioeconomic status on therapeutic clinical trial screening and enrollment. *J. Neuro-Oncol.* **2020**, *148*, 131–139. [CrossRef]
20. Ragavan, M.; Patel, M.I. The evolving landscape of sex-based differences in lung cancer: A distinct disease in women. *Eur. Respir. Rev.* **2022**, *31*, 210100. [CrossRef]
21. Mazul, A.L.; Naik, A.N.; Zhan, K.Y.; Stepan, K.O.; Old, M.O.; Kang, S.Y.; Nakken, E.R.; Puram, S.V. Gender and race interact to influence survival disparities in head and neck cancer. *Oral Oncol.* **2021**, *112*, 105093. [CrossRef]
22. Carrano, A.; Juarez, J.; Incontri, D.; Ibarra, A.; Cazares, H.G. Sex-Specific Differences in Glioblastoma. *Cells* **2021**, *10*, 1783. [CrossRef]
23. Massey, S.C.; Whitmire, P.; Doyle, T.E.; Ippolito, J.E.; Mrugala, M.M.; Hu, L.S.; Canoll, P.; Anderson, A.R.; Wilson, M.A.; Fitzpatrick, S.M.; et al. Sex differences in health and disease: A review of biological sex differences relevant to cancer with a spotlight on glioma. *Cancer Lett.* **2020**, *498*, 178–187. [CrossRef] [PubMed]
24. Tucker-Seeley, R.D. Social Determinants of Health and Disparities in Cancer Care for Black People in the United States. *JCO Oncol. Pract.* **2021**, *17*, 261–263. [CrossRef] [PubMed]
25. Leech, M.M.; Weiss, J.E.; Markey, C.; Loehrer, A.P. Influence of Race, Insurance, Rurality, and Socioeconomic Status on Equity of Lung and Colorectal Cancer Care. *Ann. Surg. Oncol.* **2022**, *29*, 3630–3639. [CrossRef] [PubMed]
26. Li, S.; Natwick, T.; Liu, J.; Morrison, V.A.; Vidito, S.; Werther, W.; Yusuf, A.A.; Usmani, S.Z. Mortality by a proxy performance status as defined by a claims-based measure for disability status in older patients with newly diagnosed multiple myeloma in the United States. *J. Geriatr. Oncol.* **2019**, *10*, 490–496. [CrossRef]
27. Al Feghali, K.A.; Buszek, S.M.; Elhalawani, H.; Chevli, N.; Allen, P.K.; Chung, C. Real-world evaluation of the impact of radiotherapy and chemotherapy in elderly patients with glioblastoma based on age and performance status. *Neuro-Oncol. Pract.* **2020**, *8*, 199–208. [CrossRef]
28. Jung, H.; Lu, M.; Quan, M.L.; Cheung, W.Y.; Kong, S.; Lupichuk, S.; Feng, Y.; Xu, Y. New method for determining breast cancer recurrence-free survival using routinely collected real-world health data. *BMC Cancer* **2022**, *22*, 281. [CrossRef]
29. Rossi, L.A.; Melstrom, L.G.; Fong, Y.; Sun, V. Predicting post-discharge cancer surgery complications via telemonitoring of patient-reported outcomes and patient-generated health data. *J. Surg. Oncol.* **2021**, *123*, 1345–1352. [CrossRef]
30. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
31. Höhn, J.; Hekler, A.; Krieghoff-Henning, E.; Kather, J.N.; Utikal, J.S.; Meier, F.; Gellrich, F.F.; Hauschild, A.; French, L.; Schlager, J.G.; et al. Skin Cancer Classification Using Convolutional Neural Networks with Integrated Patient Data: A Systematic Review (Preprint). *J. Med Internet Res.* **2020**, *23*, e20708. [CrossRef]
32. Wong, A.; Otles, E.; Donnelly, J.P.; Krumm, A.; McCullough, J.; DeTroyer-Cooley, O.; Pestrue, J.; Phillips, M.; Konye, J.; Penoza, C.; et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Intern. Med.* **2021**, *181*, 1065–1070. [CrossRef]
33. Vyas, D.A.; Eisenstein, L.G.; Jones, D.S. Hidden in Plain Sight—Reconsidering the Use of Race Correction in Clinical Algorithms. *New Engl. J. Med.* **2020**, *383*, 874–882. [CrossRef] [PubMed]

34. Lee, P.; Microsoft; Abernethy, A.; Shaywitz, D.; Gundlapalli, A.; Weinstein, J.; Doraiswamy, P.M.; Schulman, K.; Madhavan, S.; Verily; et al. Digital Health COVID-19 Impact Assessment: Lessons Learned and Compelling Needs. *NAM Perspect.* **2022**. [CrossRef] [PubMed]
35. Bose, P.S.W.; Syed, K.; Hagan, M.; Palta, J.; Kapoor, R.; Ghosh, P. Deep neural network models to automate incident triage in the radiation oncology incident learning system. In Proceedings of the BCB'21: Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, Gainesville, FL, USA, 1–4 August 2021; Volume 51, pp. 1–10.
36. Brown, W.E.; Sung, K.; Aleman, D.M.; Moreno-Centeno, E.; Purdie, T.; McIntosh, C.J. Guided undersampling classification for automated radiation therapy quality assurance of prostate cancer treatment. *Med Phys.* **2018**, *45*, 1306–1316. [CrossRef] [PubMed]
37. Liu, R.; Hall, L.O.; Bowyer, K.W.; Goldgof, D.B.; Gatenby, R.; Ben Ahmed, K. Synthetic minority image over-sampling technique: How to improve AUC for glioblastoma patient survival prediction. In Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, 5–8 October 2017; pp. 1357–1362. [CrossRef]
38. Suárez-García, J.G.; Hernández-López, J.M.; Moreno-Barbosa, E.; de Celis-Alonso, B. A simple model for glioma grading based on texture analysis applied to conventional brain MRI. *PLoS ONE* **2020**, *15*, e0228972. [CrossRef] [PubMed]
39. Li, J.; Tao, Y.; Cong, H.; Zhu, E.; Cai, T. Predicting liver cancers using skewed epidemiological data. *Artif. Intell. Med.* **2022**, *124*, 102234. [CrossRef]
40. Isensee, F.; Kickingereder, P.; Wick, W.; Bendszus, M.; Maier-Hein, K.H. Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge. In Proceedings of the International MICCAI Brainlesion Workshop, Granada, Spain, 16 September 2018; pp. 287–297. [CrossRef]
41. Goyal, S.; Raghunathan, A.; Jain, M.; Simhadri, H.V.; Jain, P. DROCC: Deep Robust One-Class Classification. In Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research, Online, 13–18 July 2020; pp. 3711–3721.
42. Gao, L.; Zhang, L.; Liu, C.; Wu, S. Handling imbalanced medical image data: A deep-learning-based one-class classification approach. *Artif. Intell. Med.* **2020**, *108*, 101935. [CrossRef]
43. Welch, M.L.; McIntosh, C.; McNiven, A.; Huang, S.H.; Zhang, B.-B.; Wee, L.; Traverso, A.; O'Sullivan, B.; Hoebers, F.; Dekker, A.; et al. User-controlled pipelines for feature integration and head and neck radiation therapy outcome predictions. *Phys. Medica* **2020**, *70*, 145–152. [CrossRef]
44. Leevy, J.; Khoshgoftaar, T.M.; Bauder, R.A.; Seliya, N. A survey on addressing high-class imbalance in big data. *J. Big Data* **2018**, *5*, 42. [CrossRef]
45. Nguyen, N.N.; Duong, A.T. Comparison of Two Main Approaches for Handling Imbalanced Data in Churn Prediction Problem. *J. Adv. Inf. Technol.* **2021**, *12*. [CrossRef]
46. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
47. Lin, T.-Y.; Goyal, P.; Girshick, R.B.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007.
48. Jaeger, P.F.; Kohl, S.A.A.; Bickelhaupt, S.; Isensee, F.; Kuder, T.A.; Schlemmer, H.-P.; Maier-Hein, K.H. Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection. In Proceedings of the Machine Learning for Health NeurIPS Workshop, Proceedings of Machine Learning Research; 2020; pp. 171–183. Available online: http://proceedings.mlr.press/v116/jaeger20a/jaeger20a.pdf (accessed on 1 December 2021).
49. Xiong, Y.; Ye, M.; Wu, C. Cancer Classification with a Cost-Sensitive Naive Bayes Stacking Ensemble. *Comput. Math. Methods Med.* **2021**, *2021*, 5556992. [CrossRef]
50. Shon, H.S.; Batbaatar, E.; Kim, K.O.; Cha, E.J.; Kim, K.-A. Classification of Kidney Cancer Data Using Cost-Sensitive Hybrid Deep Learning Approach. *Symmetry* **2020**, *12*, 154. [CrossRef]
51. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. *Front. Comput. Sci.* **2019**, *14*, 241–258. [CrossRef]
52. Sagi, O.; Rokach, L. Ensemble learning: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1249. [CrossRef]
53. Tang, X.; Cai, L.; Meng, Y.; Gu, C.; Yang, J.; Yang, J. A Novel Hybrid Feature Selection and Ensemble Learning Framework for Unbalanced Cancer Data Diagnosis With Transcriptome and Functional Proteomic. *IEEE Access* **2021**, *9*, 51659–51668. [CrossRef]
54. Le, D.N.T.; Le, H.X.; Ngo, L.; Ngo, H.T. Transfer learning with class-weighted and focal loss function for automatic skin cancer classification. *arXiv* **2020**, arXiv:2009.05977.
55. Wang, S.; Dong, D.; Zhang, W.; Hu, H.; Li, H.; Zhu, Y.; Zhou, J.; Shan, X.; Tian, J. Specific Borrmann classification in advanced gastric cancer by an ensemble multilayer perceptron network: A multicenter research. *Med Phys.* **2021**, *48*, 5017–5028. [CrossRef] [PubMed]
56. Chen, C. Using Random Forest to Learn Imbalanced Data. *Univ. Calif. Berkeley* **2004**, *110*, 24.
57. Zhao, D.; Liu, H.; Zheng, Y.; He, Y.; Lu, D.; Lyu, C. Whale optimized mixed kernel function of support vector machine for colorectal cancer diagnosis. *J. Biomed. Informatics* **2019**, *92*, 103124. [CrossRef]
58. Urdal, J.; Engan, K.; Kvikstad, V.; Janssen, E.A. Prognostic prediction of histopathological images by local binary patterns and RUSBoost. In Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO), Kos, Greece, 28 August–2 September 2017. [CrossRef]
59. Mirza, B.; Wang, W.; Wang, J.; Choi, H.; Chung, N.C.; Ping, P. Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes* **2019**, *10*, 87. [CrossRef]
60. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.

61. Hilario, M.; Kalousis, A. Approaches to dimensionality reduction in proteomic biomarker studies. *Briefings Bioinform.* **2007**, *9*, 102–118. [CrossRef] [PubMed]
62. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B-Methodol.* **1996**, *58*, 267–288. [CrossRef]
63. Yan, K.K.; Zhao, H.; Pang, H. A comparison of graph- and kernel-based–omics data integration algorithms for classifying complex traits. *BMC Bioinform.* **2017**, *18*, 539. [CrossRef] [PubMed]
64. Fawcett, T. An Introduction to ROC analysis. *Pattern Recogn. Lett.* **2006**, *27*, 861–874. [CrossRef]
65. Garcıa, V.; Sánchez, S.J.; Mollineda, R.A. *Exploring the Performance of Resampling Strategies for the Class Imbalance Problem*; Springer: Berlin/Heidelberg, Germany, 2010.
66. Lao, J.; Chen, Y.; Li, Z.-C.; Li, Q.; Zhang, J.; Liu, J.; Zhai, G. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Sci. Rep.* **2017**, *7*, 10353. [CrossRef]
67. Wu, A.; Li, Y.; Qi, M.; Lu, X.; Jia, Q.; Guo, F.; Dai, Z.; Liu, Y.; Chen, C.; Zhou, L.; et al. Dosiomics improves prediction of locoregional recurrence for intensity modulated radiotherapy treated head and neck cancer cases. *Oral Oncol.* **2020**, *104*, 104625. [CrossRef]