Received: 2 February 2023 | Revised: 27 October 2023 | Accepted: 30 October 2023

DOI: 10.1002/widm.1523

ADVANCED REVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY WILEY

The use of gene expression datasets in feature selection research: 20 years of inherent bias?

Bruno I. Grisci[1,2] | **Bruno César Feltes**[3] | **Joice de Faria Poloni**[3] | **Pedro H. Narloch**[1] | **Márcio Dorn**[1,4,5]

[1]Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

[2]Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada

[3]Institute of Biosciences, Federal University of Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

[4]National Institute of Science and Technology - Forensic Science, Porto Alegre, Rio Grande do Sul, Brazil

[5]Center for Biotechnology, Federal University of Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

Correspondence
Márcio Dorn, Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, RS, 91501-970, Brazil.
Email: mdorn@inf.ufrgs.br

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Numbers: 151591/2022-9, 314082/2021-2, 408154/2022-5, 440279/2022-4, 465450/2014-8; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Finance Code 001, Grant/Award Numbers: 88881.198766/2018-01, 88881.522073/2020-01; Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul, Grant/Award Numbers: 17/2551-0000520-1, 19/2551-0001906-8; Global Affairs Canada, Emerging Leaders in the Americas Program Scholarship

Edited by: Justin Wang, Associate Editor and Witold Pedrycz, Editor-in-Chief

Abstract

Feature selection algorithms are frequently employed in preprocessing machine learning pipelines applied to biological data to identify relevant features. The use of feature selection in gene expression studies began at the end of the 1990s with the analysis of human cancer microarray datasets. Since then, gene expression technology has been perfected, the Human Genome Project has been completed, new microarray platforms have been created and discontinued, and RNA-seq has gradually replaced microarrays. However, most feature selection methods in the last two decades were designed, evaluated, and validated on the same datasets from the microarray technology's infancy. In this review of over 1200 publications regarding feature selection and gene expression, published between 2010 and 2020, we found that 57% of the publications used at least one outdated dataset, 23% used only outdated data, and 32% did not cite data sources. Other issues include referencing databases that are no longer available, the slow adoption of RNA-seq datasets, and bias toward human cancer data, even for methods designed for a broader scope. In the most popular datasets, some being 23 years old, mislabeled samples, experimental biases, distribution shifts, and the absence of classification challenges are common. These problems are more predominant in publications with computer science backgrounds compared to publications from biology and can lead to inaccurate and misleading biological results.

This article is categorized under:

Algorithmic Development > Biological Data Mining

Technologies > Machine Learning

KEYWORDS

feature selection, gene expression, machine learning, microarray, RNA-seq

Bruno I. Grisci, Bruno César Feltes, and Joice de Faria Poloni contributed equally to this study.

© 2023 Wiley Periodicals LLC.

1 | INTRODUCTION

Feature selection comprises a wide array of algorithms and methods employed to create, among other uses, more representative datasets by filtering noisy, irrelevant, or redundant samples, leading to optimized machine learning training (Ang et al., 2016; Lazar et al., 2012; Tadist et al., 2019). As part of a knowledge discovery pipeline, feature selection is generally used to identify which features in the original data can convey relevant information. For example, feature selection was used to identify hub genes of hepatocellular carcinoma (Li & Xu, 2019), predict bioluminescent proteins (Kandaswamy et al., 2011), cluster single-cell data (Ranjan et al., 2021), and as part of a pipeline for determining battery capacity fade (Roman et al., 2021).

Because of this characteristic, feature selection became one of the most used approaches in Bioinformatics, frequently employed to process high-dimensional gene expression data (Ang et al., 2016). Gene selection is the name given to the application of feature selection to transcriptomic data, allowing the discovery of relevant expressed genes capable of separating samples from different populations or target annotations (i.e., the samples classes) (Lazar et al., 2012). Relevant genes that can satisfactorily classify a given condition are sometimes referred to as informative genes, which could be used in diagnosing diseases or as potential drug targets (Lazar et al., 2012). Therefore, in recent years, feature selection has been discussed as a tool for uncovering potential tumoral biomarkers, allowing reliable diagnosis and prognosis of different cancer types (Grisci et al., 2018, 2019). Numerous works provide complete reviews of feature selection algorithms and their application to gene expression data (Ang et al., 2016; Bolón-Canedo et al., 2014; Boulesteix et al., 2008; Feltes et al., 2018; Lazar et al., 2012; Osama et al., 2022; Saeys et al., 2007). According to a survey by Osama et al. (2022), between 2010 and 2021, the number of publications on gene selection increased by 1.8-fold, and the citations by 135.5-fold.

However, discussing how and why feature selection can be applied to gene expression data goes beyond which algorithms should be employed. Because the accuracy, performance, and final results of any feature selection algorithm depend on the nature and quality of the initial input, it is indispensable to explain how gene expression data were developed and analyzed over time—a topic absent from most discussions in the field.

Large-scale gene expression techniques emerged in the mid-90s and started even before the Human Genome Project (HGP), which was finished in 2003, covering about 92% of the human genome (Figure 1). Only in 2022 did new sequencing technologies providing long-read sequences allow the completion of the human genome (Nurk et al., 2022). Since the development of microarray technology, concerns regarding poor reproducibility between different microarray platforms have emerged. One of the reasons is related to the chosen probes and probe sets (Liu et al., 2010). Probes are designed to hybridize to a messenger RNA (mRNA) molecule based on expressed sequence tag, complementary DNA (cDNA), or mRNA deposited in the NCBI repository (Liu et al., 2010). In the early years of microarray, it was not unusual to find probes prone to non-specific hybridization, especially in non-well-documented organisms, where the source sequence used to design the probe was more likely to have inaccurate or incomplete annotations, presence of sequencing artifacts, and redundant sequences (Liu et al., 2010). For example, 30%–40% of probes from Affymetrix GeneChip (Figure 1), the most popular microarray platform, showed discrepancy with gene and transcript definitions (Dai et al., 2005; Gautier, Møller, et al., 2004), where more than 5000 probes presented cross-hybridization issues due to splice variants or closely related genes (Harbig et al., 2005). Compared to data available today, older chips could be composed of 5000–8000 probes, whereas the modern, most frequently used ones can reach between 22,000 and 70,000 probes, depending on platform and manufacturer. Likewise, many of those probes were discontinued due to their inaccuracy.

As seen in Figure 1, there is a 9-year gap between the creation of Affymetrix, the most used microarray platform, and the end of the HGP. The computational analyses of microarray data started at the end of the 1990s, and feature selection became one of the standard methods of data investigation in the field (Bolón-Canedo et al., 2014; Saeys et al., 2007). The first significant application of feature selection and machine learning to gene expression data was pioneered in 1999 by Alon et al. (1999) and Golub et al. (1999); hence, 4 years before the HGP finishing and the release of the official version of the human genome, 3 years from the creation of the Gene Expression Omnibus (GEO) database (Edgar et al., 2002), the largest gene expression database, and 2 years before the release of the initial draft of the human genome. At the time, the successful application of the classification task to gene expression data was a great leap that invariably opened the door for what is now one of the most fertile grounds for computational biology applied to omics data. They were followed by Bhattacharjee et al. (2001) and Khan et al. (2001) in 2001, and Shipp et al. (2002) in 2002 (Mramor et al., 2007). Despite all of these authors focusing on the study of distinct types of cancer, this shared object of research settled a clear bias toward the selection of human cancer datasets for the experiments of later feature selection research.

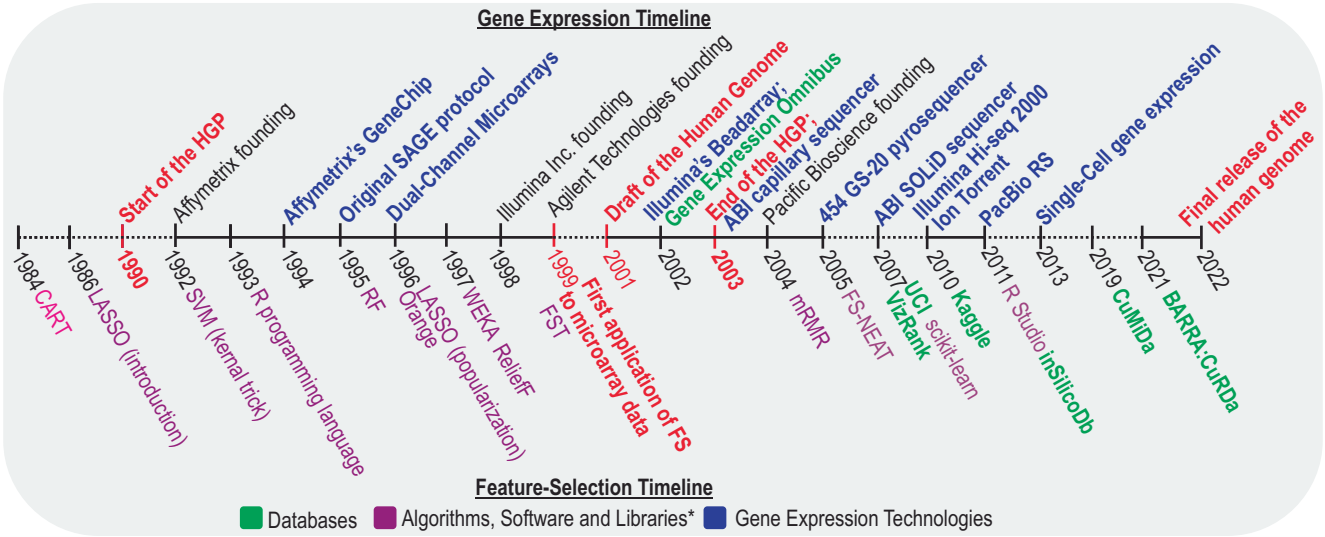

FIGURE 1 Timeline of significant impacts on gene expression and feature selection fields. Dashed lines indicate the passage of multiple years. The stepping stones are: Classification And Regression Trees (CART) (Breiman et al., 2017); Least Absolute Shrinkage and Selection Operator (Lasso) (Santosa & Symes, 1986; Tibshirani, 1996); Human Genome Project (HGP); Support Vector Machine (SVM) (Boser et al., 1992); Random Forest (RF) (Ho, 1995); Serial Analysis of Gene Expression (SAGE); ReliefF (Robnik-Šikonja & Kononenko, 1997); Waikato Environment for Knowledge Analysis (WEKA) (Holmes et al., 1994); Feature Selection Toolbox (FST); minimum Redundancy—maximum Relevance (mRMR) (Peng et al., 2005); Feature Selective Neuroevolution of Augmenting Topologies (FS-NEAT) (Whiteson et al., 2005); VizRank (Leban et al., 2006); UC Irvine Machine Learning Repository (UCI) (Dua & Graff, 2017); Scikit-learn (Pedregosa et al., 2011); inSilico database (inSilicoDb) (Taminau et al., 2011); Curated Microarray Database (CuMiDa) (Feltes et al., 2019); Benchmarking of ARtificial intelligence Research: Curated RNA-seq Database (BARRA:CuRDa) (Feltes et al., 2021). Asterisk indicates algorithms, software, and libraries commonly used or created for feature selection or hallmarks in the field.

Also, during the 2000s, innovations in data analysis and feature selection were being developed (Mramor et al., 2007), where the univariate paradigm was the most popular type of feature selection algorithm. These algorithms were fast and scaled well with the increased size of datasets but did not account for feature dependencies (Bolón-Canedo et al., 2014). Consequentially, other methods, such as wrappers or statistical tools, were soon adapted or developed to deal with the new technology (Bolón-Canedo et al., 2014; Mramor et al., 2007; Saeys et al., 2007). However, as gene expression technology advanced, with new manufacturers being inserted into the market and new microarray platforms being continuously created—and discontinued—most algorithms being developed did not follow. They kept being designed, evaluated, and validated with the same datasets from the days of the microarray technology infancy in mind.

As it is standard in the Computational Sciences, the classification success of new algorithms must be compared to what was previously published. Thus, some earlier datasets (fully discussed in Section 4), were used as inputs, and their results as the basis for "successful" accuracy results. Unfortunately, while new platforms kept being updated with the continuous genetic knowledge being published and deposited in databases and new datasets repeatedly made available at GEO, the same old datasets kept being solely employed to train and test new algorithms. The culture of reusing the same datasets did not change, even after RNA-seq became popular around 2010 (Figure 1).

Errors and biases in popular datasets used for analyses and benchmarks in machine learning and bioinformatics are not unheard of (Liang et al., 2022; Nature, 2022). For example, Northcutt et al. (2021) identified that, on average, around 3.4% of the labels in the 10 most used datasets of computer vision, natural language, and audio contain errors. The authors discuss that because such datasets are used to validate findings or to measure the state-of-the-art, even a small percentage of data errors can lead to incorrect conclusions about the performance of new methods. For instance, a low increment in the accuracy of a dataset could be coming from the model overfitting the wrong labels present in the dataset. These errors and biases can also impact computational methods' later usefulness or safety in real-world applications. In a systematic review of machine learning models trained for the detection and prognostication of coronavirus disease 2019 (COVID-19) from chest radiography or chest computed tomography images, Roberts et al. (2021) discovered that none could be used in clinical practice because of methodological flaws or underlying biases. Among the reported issues was bias due to the small sampling size of patients with COVID-19 when considering the variability

of large international datasets. Likewise, benchmarking studies focusing on employing fewer datasets to evaluate tasks for which the data was not initially designed are also current issues in machine learning (Koch et al., 2021). As discussed in Sections 2 and 3, this scenario is not distant from the current application of feature selection to gene expression data analysis.

The remainder of the text is organized as follows. Section 2 describes how gene expression data (microarray and RNA-seq) is seen from a biological and computational perspective, and discusses how researchers from both fields usually handle them. Section 3 reviews the feature selection literature to identify the most commonly used datasets and their characteristics. Section 4 discusses some of the issues of these commonly used datasets and how they can impact the results of new experiments. Section 5 is a comment on the challenges of creating or finding reliable databases to share or access gene expression data for computational research and a brief review of the main online databases. Section 6 concludes this study with perspectives for the field and recommendations for researchers and reviewers based on our findings.

2 | DATA BETWEEN WORLDS

Computationally speaking, gene expression data is usually represented as a matrix of continuous numerical values (Whitworth, 2010), where each row represents a sample, and each column represents a gene. In this model, a matrix cell contains values that measure the gene expression level in that sample. Usually, the samples are categorized in biological conditions, such as the division between healthy and tumor samples, which will be compared. Gene expression studies centered on the biological aspect of a given research question (i.e., not focused on developing a new algorithm), being either purely wet-lab or dry-lab-derived data, will generally employ known computational tools to observe differentially expressed genes (DEG). Bioinformatic analyses without creating a new algorithm are far from straightforward or lacking complexity. There are dozens of known protocols, workflows, combined approaches, web tools, standalone software, and R packages available for such tasks, which are also distinct between microarray and RNA-seq.

Microarrays are experiments in which predefined probes will hybridize with a biotinylated cDNA molecule and generate fluorescent light emission that can be detected by a scanner and quantified by the machine in log-intensities values (Blalock, 2003; Epstein & Butow, 2000). A multitude of software can then analyze the raw data. For example, different microarray platform manufacturers such as Affymetrix, Illumina, and Agilent have their own R packages dedicated to analyzing their own generated datasets, such as *affy* (Gautier, Cope, et al., 2004) and *oligo* (Carvalho & Irizarry, 2010) for Affymetrix, *illuminaio* (Smith et al., 2013), *beadarray* (Dunning et al., 2007), and *lumi* (Du et al., 2008) for Illumina, and *agilp* (Chain, 2021) for Agilent. In addition, *limma* (Ritchie et al., 2015) is also one of the most employed R packages, with a wide range of functions for all types of platforms, especially for extracting DEG. On top of such tools, other packages provide additional analyses and refinement options for gene expression studies, such as the *arrayQualityMetrics* package (Kauffmann et al., 2008), which assesses sample quality of microarray datasets, and *biobase* (Huber et al., 2015), which provides a myriad of information for different platforms, probes, and genes that greatly aids the analysis process. Numerous other options can be found on the Bioconductor (bioconductor.org) database.

RNA-seq significantly differs from microarrays in both technical aspects and results. Contrary to microarrays, in which probes need to be previously designed to detect a given gene expression, RNA-seq allows the quantification of the total RNA expression profiling in a single experiment. This technique offers the identification of novel transcripts (de novo transcriptome assembly), allele variants (as long as they are present in the expressed genes), and analysis of alternative splicing variants. In addition, RNA-seq reaches lower technical variability than microarray, and it is concordant with other transcriptomic techniques, such as qRT-PCR (Corchete et al., 2020). With so many applications, each scenario must be carefully evaluated to create the appropriate experimental design and bioinformatics workflow according to the organism and the research goals. For example, to perform microRNA (miRNAs) (RNAs with 21–25 nucleotides in length) sequencing, total RNA fractionation and enrichment of miRNAs by size must be conducted during library preparation; otherwise, they will be leached out due to their small size. Furthermore, RNA-seq can be used as the unique transcriptome profiling method or combined with additional assays, such as CITE-seq (Stoeckius et al., 2017) and REAP-seq (Peterson et al., 2017).

RNA-seq presents detailed steps, including RNA extraction and purification, library construction, sequencing, and bioinformatics analysis. Each error or bias introduced in any stage can interfere with the next step, directly affecting the sequencing quality and interpretation problems (Shi et al., 2021). The step with the highest number of biases is

library preparation, which should receive additional attention as it can strongly affect the quality of the final data (Shi et al., 2021). RNA-seq analysis of downstream sequencing is dependent on technology application. Still, the primary analysis step includes quality control, read mapping (using a reference genome/transcriptome or de novo assembly) to infer the expressed transcripts, and abundance estimation based on the number of mapped reads in the sequence region of a transcript or gene. However, some tools perform alignment-independent transcript quantification (Conesa et al., 2016). Thus, RNA-seq matrices are not based on log-intensity values but are on read counts.

Habitually, biologically focused works will employ classical analysis protocols or innovate how the data is analyzed by creating new pipelines that increase the robustness of the results or by developing integrative pipelines to validate the results using multiple computational approaches. Independently of the chosen workflow, gene expression microarray analyses that aim to identify DEG will perform: (i) a background correction, which aims to remove artifacts from the raw data; (ii) a normalization step. In this case, the normalization can be performed within the array and between arrays, depending on the platform type; (iii) quality analysis, which will access the sample quality of the dataset. Although highly recommended, this step is not employed by all researchers; (iv) identification of the DEG. On the other hand, RNA-seq studies will usually have: (i) a normalization, where it is important to consider that different samples may have different library sizes, which means that samples with larger library sizes will show more reads mapped to each gene. Also, the gene length must be considered since longer genes will have more reads mapped to them. In this sense, RNA-seq normalization is crucial for downstream analysis and should always consider library size and gene length; (ii) batch effects correction, which could be minimized during experimental design but can be considered by batch correction methods; (iii) identification of DEG. The steps mentioned above are all included in some popular methods, such as *edgeR* and *DESeq2*, which consider raw reads as input (Conesa et al., 2016; Love et al., 2014; Robinson et al., 2010).

Remembering that a classical biological approach focuses on finding biological novelty independently of the employed protocol is essential. Likewise, creating new microarray and RNA-seq datasets is limited by budget and experimental conditions. Consequentially, it is not guaranteed that a massive amount of samples will be analyzed either because they might be excluded due to contamination or because of the challenges associated with acquiring them, such as patient availability and agreement or difficult access (e.g., brain samples). Therefore, studies are organized to have at least three control and three experimental samples, the standard in the field for proper statistical analysis.

Meanwhile, researchers from the Computational Sciences may be concerned about more practical aspects of the data. For instance, Ang et al. (2016) and Bolón-Canedo et al. (2014) highlight several of the challenges related to gene expression data on feature selection and machine learning research, among them the curse of dimensionality, small sample size (often less than a hundred), mislabeled data, imbalanced classes, presence of outliers, data complexity, shifts in the data distribution, and cross-platform comparisons. In the latter, comparing data from different technologies may introduce biases that make the combination of several datasets impossible due to the differences in the organization, presentation, and values between different manufacturers. Likewise, combining different experimental conditions should be made with caution because not all biological conditions can be safely compared. More issues are discussed in Section 4. These challenges are not new, and several of them have been discussed since the creation of the microarray technology (Allison et al., 2006; Leung & Cavalieri, 2003). For practical reasons, these issues may become the center of attention in computational studies, and the biological aspects end up in the background regarding the choice of datasets for experiments. However, standard practices to mitigate some of these challenges, such as class imbalance control and cross-validation, cannot overcome problems in data collection (Sambasivan et al., 2021). Thus, the mindset of both major fields involved in creating and analyzing gene expression data diverges in multiple aspects that should be considered when discussing handling datasets and applying distinct algorithms and protocols.

3 | TO TEACH A NEW MACHINE OLD TRICKS

As discussed in the previous section, there is a difference in priorities when choosing which datasets to use in experiments regarding Biological or Computational Sciences. Due to the several practical challenges these datasets offer, researchers proposing or analyzing new feature selection methods tend to stick to the already popular ones. This effect is magnified by the available online databases (Section 5) and by the lack of standard state-of-the-art results to achieve fair comparative analyses (Bolón-Canedo et al., 2014). In this case, researchers usually keep using the same datasets employed in previous publications to conduct metric comparisons. In this sense, the choice of datasets based on the most reported cases in the literature leads to a cycle; the older and known datasets are used by newer studies solely

because they are frequently cited (Beker et al., 2022), not because they are relevant, thus, placing the quality or relevance of the data in second place.

According to their review of feature selection published between the years 2011 and 2016, Ang et al. (2016) identified the five most commonly used microarray datasets to be: (i) colon cancer by Alon et al. (1999) in 1999; (ii) leukemia by Golub et al. (1999) in 1999; (iii) diffuse large B-cell lymphoma (DLBCL), from Alizadeh et al. (2000) in 2000; (iv) small round blue cell tumor (SRBCT) of childhood from Khan et al. (2001) in 2001; and (v) prostate cancer by Singh et al. (2002) in 2002. By comparing the year in which they were reviewed by Ang et al. (2016) (2011–2016) and the years they were published (1999–2002), it is possible to note that the age gap was between 14 and 17 years.

An earlier review of microarray datasets and feature selection conducted by Bolón-Canedo et al. (2014) in 2014 pinpointed some of the most popular microarray datasets used for experiments, but, unfortunately, the exact methodology employed to select publications is not clear from the study. Their results compiled 64 datasets, even though some are alternative versions of the same original data (e.g., some datasets are split into training and test sets). Alarmingly, the original reference for three of these datasets is considered unknown. All 5 datasets identified by Ang et al. (2016) are present in this list of 64 datasets by Bolón-Canedo et al. (2014). It is also worth noting that most datasets listed by Bolón-Canedo et al. (2014) are related to human cancers. Although Bolón-Canedo et al. (2014) published their results in 2014, the average year of creation of the listed datasets is 2002, once again showing a selection bias toward datasets from the late 1990s and early 2000s.

We conducted an extensive literature review to verify the existence of a significant age gap between new gene expression research and the datasets used in feature selection experiments. Our analysis resulted in a curated list of 1284 papers on feature selection applied to gene expression data, published between 2010 and 2020. We manually extracted the information from these papers of which microarray or RNA-seq datasets were being used in the experiments. As can be seen in Figure 2, during the selected period, the publication of feature selection research related to gene expression had a steady growth, while the amount of gene expression data deposited in GEO experienced exponential growth, signaling that authors had a significant increase in dataset options to choose from but chose older datasets to train and validate their algorithms. Hence, any bias regarding the age of the employed datasets cannot be attributed to data shortage.

Published works from the last 10 years on the topic of feature selection applied to gene expression data were gathered from different databases, namely: (i) Pubmed (pubmed.ncbi.nlm.nih.gov/); (ii) IEEE Xplore (ieeexplore.ieee.org); (iii) SCOPUS (scopus.com/home.uri); and (iv) Web of Science (clarivate.com/products/web-of-science/). Only works written in English from January 2010 up to December 2020 were considered. To download all possible relevant works, we only selected studies that contained the keywords "feature selection" together with either the keywords "microarray" or "RNA-seq"—thus, considering a wide range of publications. We regarded full publications published in journals and conferences; however, book chapters, abstracts, expanded abstracts, posters, and review articles were omitted. Publications that were not peer-reviewed or in preprint formats at the time of this search were not considered.

Each study was then manually, one by one, accessed to: (i) identify which and how many datasets were employed to train/test the proposed methods; (ii) observe if the authors cited the publication of the original dataset or if they cited an intermediate work that, in its turn, cited the original dataset; (iii) see which type of gene expression dataset the authors tested their approach (microarray or RNA-seq); (iv) determine if the author's employed simulated datasets; (v) recover the years the datasets were made available; (vi) identify the biological background (leukemia, breast cancer, hepatitis, among others) and organism, such as *Homo sapiens* or *Mus musculus*, of each dataset; (vii) inspect if datasets cited only using URLs were still available online. Datasets in which the URL led to a dead-end were excluded; (viii) investigate if the datasets that were only cited using URLs were easily obtainable. For example, if they are only redirected to a general website and the dataset's location is unclear. Studies in which the URLs lead to a non-specific section of a website and the datasets could not be located were excluded. All publications and dataset information evaluated in our work is available in Tables S1–S3 in the Supplementary Material 2.

An overview of the collected data is shown in Table 1. Our results show that, on average, each article uses around four distinct datasets. Among them, 32.3% of dataset citations did not cite the original study—instead, they referenced intermediate publications, which we classified as "indirect citation." This was prevalent in papers from repositories focused on the Computational Science and engineering fields (39.4% of papers from IEEE Xplore) and less common in articles from biology-focused repositories (15.5% of papers from PubMed). The same pattern appears for articles with broken links to their data sources or studies missing the data reference altogether. This indicates a lack of concern regarding the origin and access of gene expression data used in around one third of the publications on feature selection, especially the research from the Computational Science field. Not only does this shed light on the fact that some

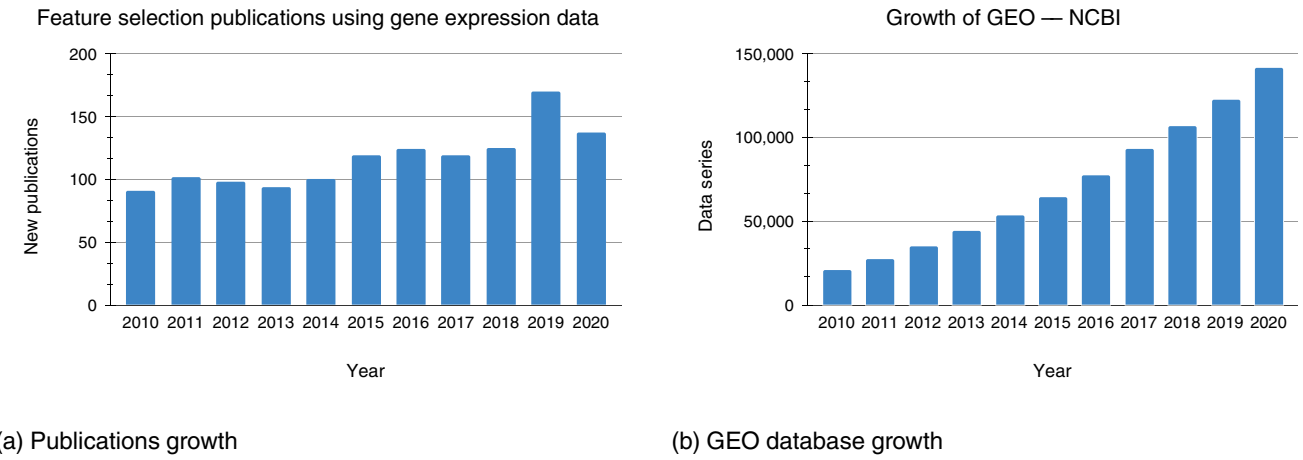

FIGURE 2 Comparison between the growth in feature selection applied to gene expression studies and the new datasets deposited in the GEO database. Both charts cover the period between 2010 and 2020.

TABLE 1 Summary statistics of the publications reviewed in this study.

	All	IEEE Xplore	PubMed	WOS	Scopus
Number of publications	1284	375	315	47	547
Mean number of datasets per publication	3.94 ± 3.38	3.76 ± 3.17	3.66 ± 3.47	3.66 ± 3.03	4.26 ± 3.48
Median number of datasets per publication	3	3	2	3	3
Min number of datasets per publication	1	1	1	1	1
Max number of datasets per publication	27	27	25	12	27
Indirect citations[a]	32.3%	39.4%	15.5%	23.4%	37.8%
Broken links[b]	7.7%	8.5%	5.3%	10.6%	8.4%
Missing reference[c]	5.1%	9.6%	0.3%	6.3%	4.7%
Simulated data[d]	6.5%	8.0%	8.5%	6.3%	4.3%
Author's dataset[e]	4.6%	2.1%	9.2%	4.2%	3.8%

[a]Articles citing an intermediary study instead of the original source of the data.
[b]Articles containing URLs that do not work or are not up-to-date.
[c]Articles without proper references for the used datasets.
[d]Articles using data from computational simulations and not from biological experiments.
[e]Articles using their own datasets.

feature selection experiments may be conducted on outdated data, but it also damages the reproducibility of the findings, as the information on the original datasets is missing.

The most commonly used datasets (appearing in over 10% of the publications) found in our research are listed in Table 2. It shows that 57% of the publications used at least one of the datasets in Table 2, and 23% of them only used datasets from this list. The top five datasets in Table 2 were used by 53% of the reviewed publications. The leukemia dataset from Golub et al. (1999) is so prominent that it was employed by 40% of the publications and around 2% of the publications used only this single dataset in their experiments. These popular datasets all have some characteristics in common: they are microarray experiments from the beginning of the popularization of gene expression analysis (between 1999 and 2002), all related to human cancers, and with a small number of features for today's standards (less than 25,000). Thus, it is somewhat alarming that these older datasets are still widely used in the analyzed period. When compared with the timeline in Figure 1, it is clear that these datasets predate several advancements in gene expression technology and data repositories. As shown in Figure 3, the use of these 11 datasets did not decline and remained stable from 2010 to 2020, following the total number of publications (Figure 2a). Details and issues of these datasets are discussed in Section 4.

TABLE 2 The most common datasets used in experiments of the reviewed publications in the period of 2010–2020.

Ranking	Dataset	Year	Samples	Features	Classes	Background	Prevalence in the reviewed publications (%)
1	Golub et al. (1999)	1999	72	7129	2	Leukemia	40.26
2	Alon et al. (1999)	1999	62	2000	2	Colon cancer	32.13
3	Singh et al. (2002)	2002	136	12,600	2	Prostate cancer	22.98
4	Pomeroy et al. (2002)	2002	90	5920	5	Brain cancer	17.67
5	Khan et al. (2001)	2001	83	2309	4	SRBCT	15.94
6	Shipp et al. (2002)	2002	77	7129	2	DLBCL	14.54
7	Alizadeh et al. (2000)	2000	96	4026	9	Lymphoma	13.60
8	Gordon et al. (2002)	2002	181	12,533	2	Lung cancer	13.52
9	Armstrong et al. (2002)	2002	72	11,225	3	Leukemia	12.58
10	Van't Veer et al. (2002)	2002	97	24,481	2	Breast cancer	11.18
11	Bhattacharjee et al. (2001)	2001	203	12,601	5	Lung cancer	10.39

Abbreviations: DLBCL, diffuse large B-cell lymphoma; SRBCT, small round blue cell tumor.

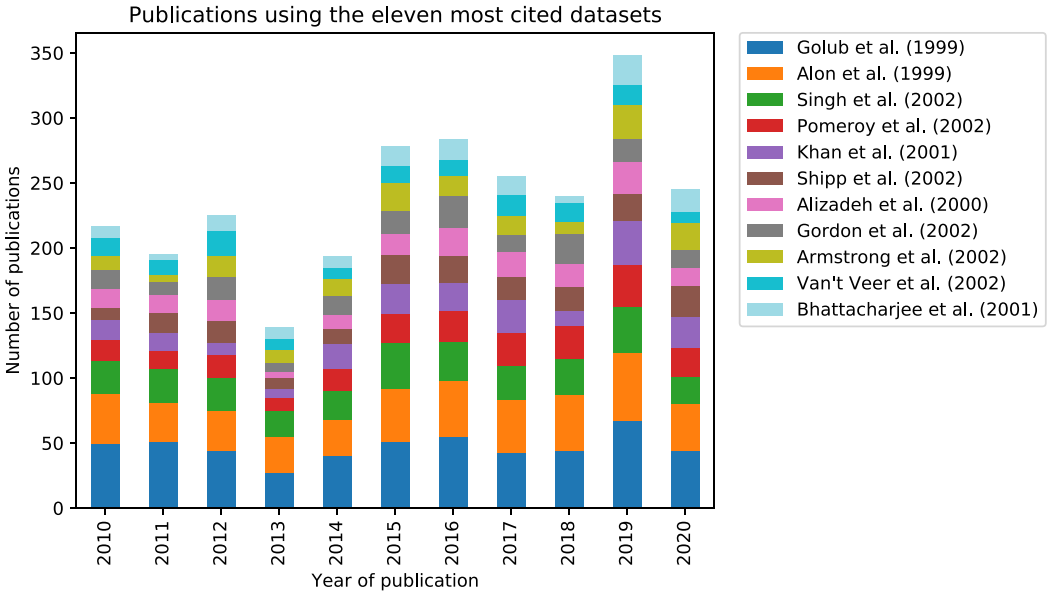

FIGURE 3 Yearly quantity of reviewed publications using each of the most common datasets from Table 2. The sum of each bar is potentially larger than the total number of publications for that year because the same study might employ one or more datasets.

Other frequently used datasets found in our review are listed in decreasing order of use in feature selection literature: Nutt et al. (2003) (brain cancer), Ramaswamy et al. (2001) (multiple cancers), Beer et al. (2002) (lung cancer), Su et al. (2002) (multiple tissues), Yeoh et al. (2002) (leukemia), West et al. (2001) (breast cancer), Staunton et al. (2001) (multiple cancers), Ross et al. (2000) (leukemia), Rosenwald et al. (2002) (lymphoma), Wigle et al. (2002) (lung cancer), Notterman et al. (2001) (colorectal cancer), Haslinger et al. (2004) (leukemia), Welsh et al. (2001) (prostate cancer), Stienstra et al. (2010) (hepatitis), Dyrskjøt et al. (2003) (bladder cancer), Hedenfalk et al. (2001) (breast cancer), Chiaretti et al. (2004) (leukemia), and Chowdary et al. (2006) (breast cancer). These microarray datasets were published between 2000 and 2006 and focused on human cancers. The only exception is Stienstra et al. (2010), a study from 2010 regarding hepatitis in *Mus musculus*. The complete list of datasets is available in Table S2 in the Supplementary Material 2.

The tendency toward experiments being conducted on older datasets was present for most of the research publications we reviewed and did not improve considerably between 2010 and 2020 (Figure 4). Although the feature selection

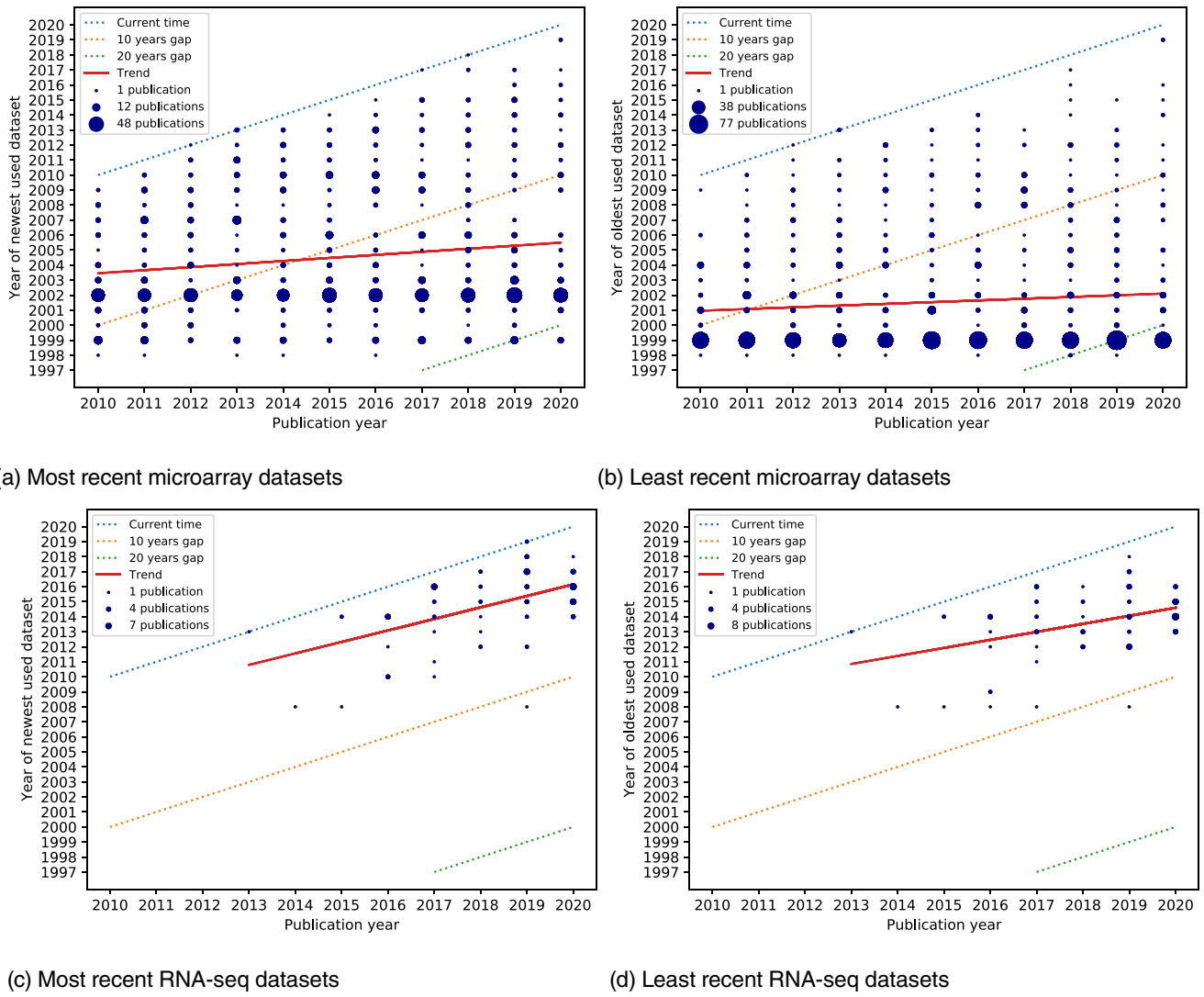

FIGURE 4 Year of publication of publications about feature selection versus the year of datasets creation. The *x*-axis displays the years of publication (2010–2020) of all reviewed literature on feature selection for gene expression data. The *y*-axis indicates the years (1997–2020) of the creation of the datasets used in experiments of the reviewed publications. Each point corresponds to all publications published in year *x* using datasets from year *y*. The size of the dot is proportional to the number of publications. The solid red line shows the trend year of the dataset creation. The dashed lines demarcate 10-year periods. (a) *y*-axis displays the year of creation of the most recent microarray dataset used in experiments; (b) *y*-axis displays the year of creation of the least recent microarray dataset used in experiments; (c) *y*-axis displays the year of creation of the most recent RNA-seq dataset used in experiments; (d) *y*-axis displays the year of creation of the least recent RNA-seq dataset used in experiments. The lines showing the trend of datasets year of creation were computed using the polyfit method (https://numpy.org/doc/stable/reference/generated/numpy.polyfit.html) from the Numpy library (Harris et al., 2020) with degree equal to one. It fits a polynomial of degree one that minimizes the squared error in the data of articles publication year versus datasets year.

literature seems to keep pace with the release of newer datasets when it comes to RNA-seq datasets, our data indicate that the vast majority of publications still heavily rely on gene expression data from older microarray experiments. If we compute a trend line of the average year the datasets were created, the trend is for the age gap issue to persist in the foreseeable future if no changes in research and editorial practices occur (Figure 4). As previously discussed, Table 1 shows that publications from the IEEE Xplore repository were more likely to have indirect citations, broken links, and missing references than publications from PubMed. Moreover, Figure 5 shows that publications from the IEEE Xplore also heavily rely on older datasets than publications from PubMed (the visualization for the WOS and Scopus databases is in Figure 1 of Supplementary Material 1). One hypothesis is that because the IEEE Xplore publications come from computer science and engineering backgrounds, their experiments are more focused on the comparison with older

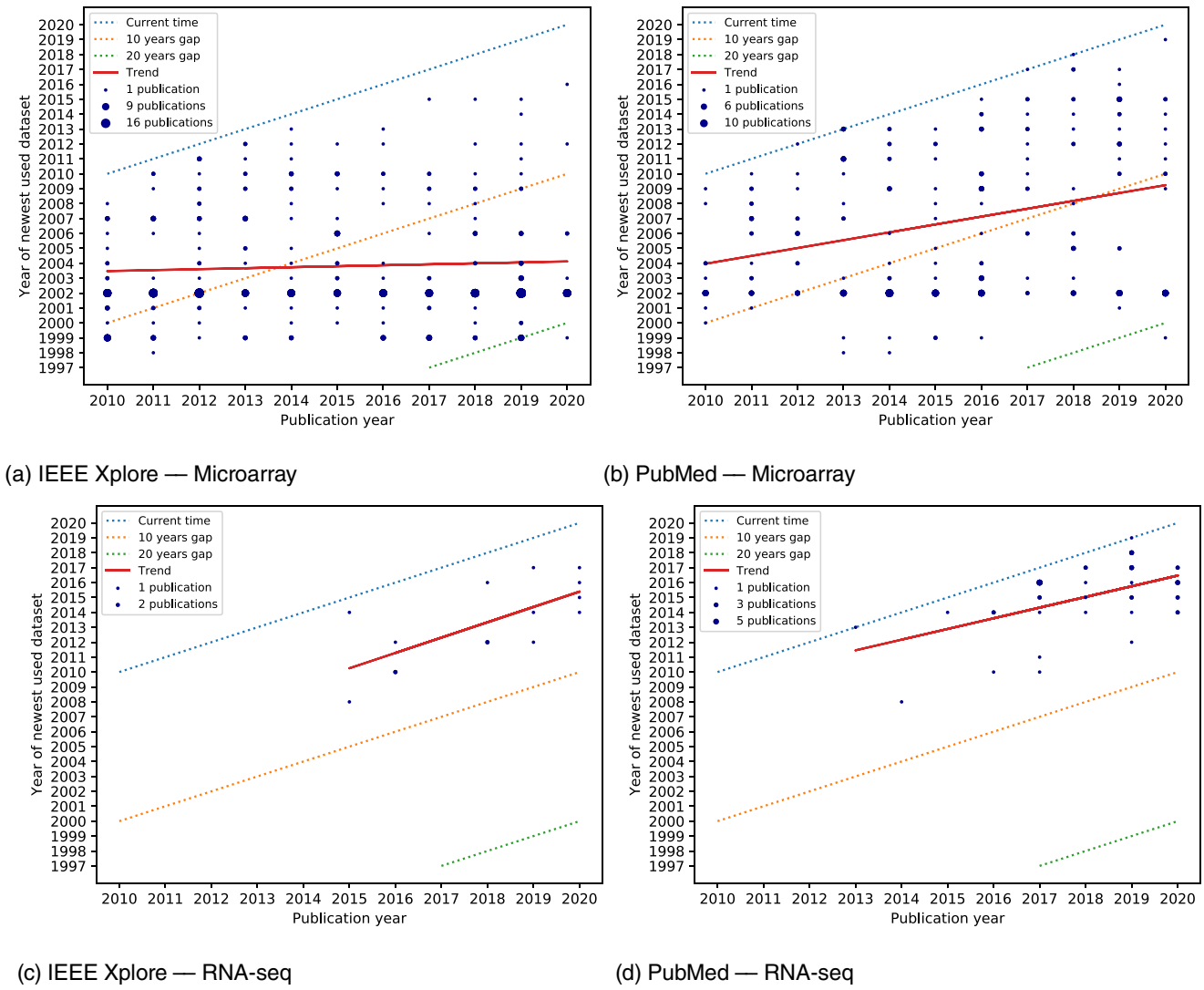

FIGURE 5 Differences between publications archived in the IEEE Xplore and PubMed databases. The chart details are as in Figure 4. (a) Year of creation of the most recent microarray dataset used in experiments of publications from IEEE Xplore; (b) year of creation of the most recent microarray dataset used in experiments of publications from PubMed; (c) year of creation of the most recent RNA-seq dataset used in experiments of publications from IEEE Xplore; (d) year of creation of the most recent RNA-seq dataset used in experiments of publications from PubMed.

methods. For this reason, the datasets used in previous publications end up shared in popular online databases, only increasing their reach, as discussed in Section 5. Meanwhile, publications from PubMed, mainly from biology-focused fields, are more concerned with the data being analyzed since it directly impacts the biological importance and accuracy of the results. Hence, studies from Biological Sciences tend to pay extra attention to data quality, the origin of the dataset, and the conditions in which it was obtained.

A final issue of the datasets employed in the past and current feature selection research is their subject of analysis. The vast majority of datasets analyzed in this work come from human cancer. As shown in Figure 6, most gene expression data employed in feature selection studies come from *Homo sapiens*, and most of the microarray datasets are related to some cancer type, which is expected due to the high relevance of the subject and its applicability to Biological and Medical Sciences. Likewise, Figure 2 in Supplementary Material 1 shows the overall distribution of data by species available on GEO, where *H. sapiens* is the most prevalent. However, it is common for algorithms developed to analyze gene expression data to be published claiming to be generalists. In this sense, the authors do not specify preconditions on which datasets to which they can apply their methods. This is a crucial factor because of the genetics and the expression patterns of *H. sapiens* and common model organisms, such as

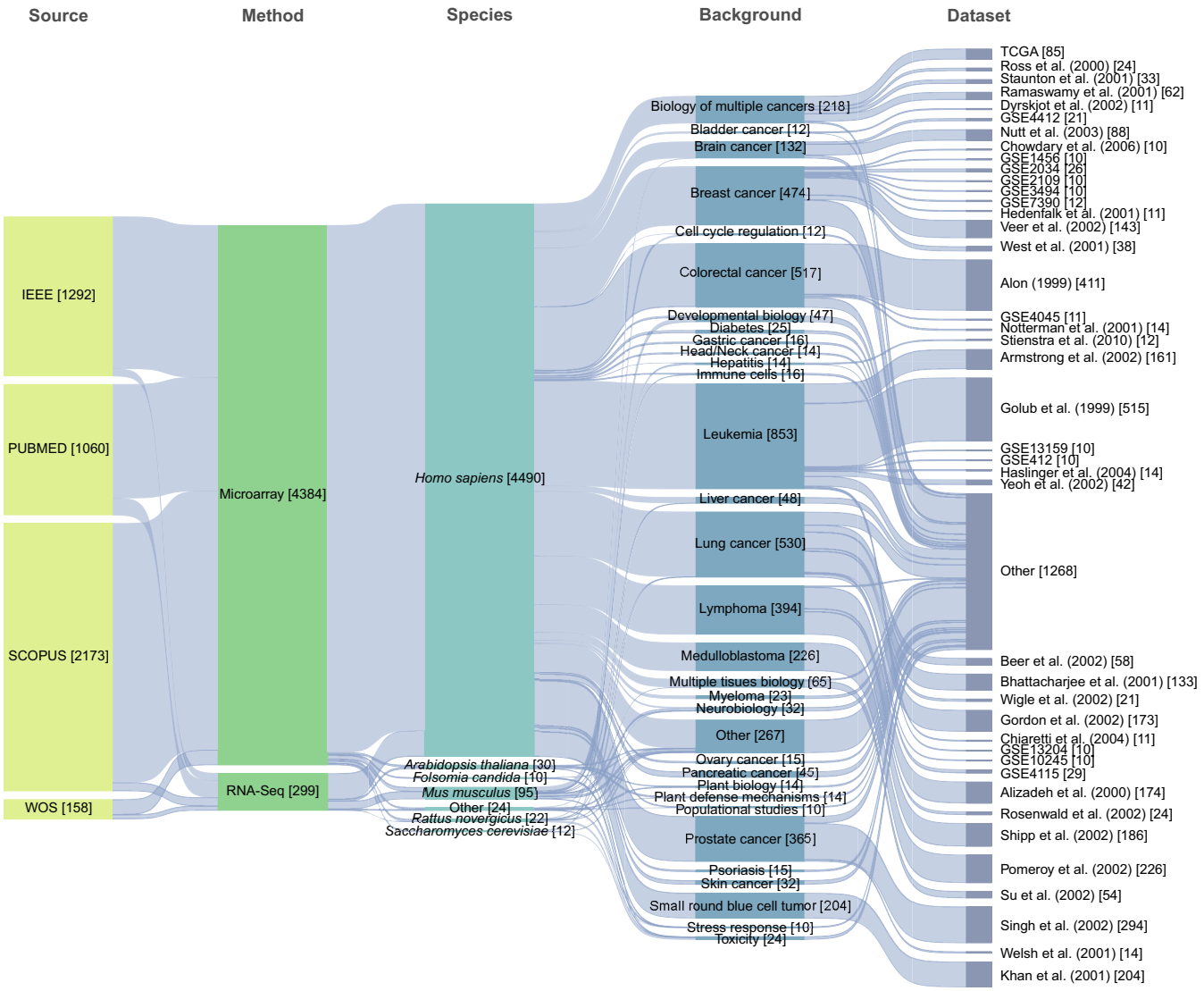

FIGURE 6 This Sankey diagram (Otto et al., 2022) was designed considering background information of the datasets employed in the reviewed publications to show the number of times the following categories appeared: (i) the origin database of each reviewed feature selection publication (the same publication can use several datasets); (ii) the transcriptomic method used by the datasets (microarray or RNA-seq); (iii) the species from where the samples were obtained; (iv) the datasets' biological background (i.e., diseases); (v) and the employed datasets. The number inside the square brackets in each category is the number of times a given information appears in the reviewed publications. The "other" in each category encompasses information that appeared equal to or less than 10 times in the reviewed publications. As discussed in the main text, some publications do not detail the employed datasets (i.e., non-defined and missing information), and those are not accounted for in this diagram.

Mus musculus, *Rattus rattus*, and *Rattus novergicus*, although similar, still have differences. Consequentially, there are variations in the experimental protocols employed in generating datasets for humans and model organisms. For example, genetic, transcriptomic, and pharmacological manipulation is common to induce tumors in rodents, which differs from studying spontaneous tumors. Hence, these datasets have different number of features, as well as numerous probes made specifically for each species. Likewise, biological backgrounds should never be overlooked since each disease has its own gene expression pattern.

In this case, the algorithms should be tested on data from several distinct organisms and biological backgrounds to avoid biases in the results. Thus, if an algorithm claims to apply to "microarray data," for instance, it must be evaluated with a representative sample of datasets that encompass their diversity, and not only on human or cancer data (Figure 6).

4 | THE DANGERS OF LIVING IN THE PAST

As discussed in the previous sections, the most overlooked aspect of feature selection applied to gene expression data was the age of the employed datasets, where a small group of dated datasets was recurrently used to train and test the algorithms, a practice that eventually became a liability as the gene expression technology advanced. In addition, errors or methodological flaws in the data creation or during usage may introduce biases that jeopardize later analyses or comparisons.

Good examples are the Singh et al. (2002) (prostate cancer) and Gordon et al. (2002) (lung cancer) datasets, which were both published in 2002 and figure among our top 10 most used datasets (third and eighth places). In the case of Singh et al. (2002), the distribution of classes in the training set (49%/51%) is different from the test set (26%/74%). Moreover, this test set has an almost 10-fold difference in microarray intensity from the training set (Bolón-Canedo et al., 2014). Such disparities in class distribution are also present in Gordon et al. (2002). This dataset has a 50%/50% split in training but a 90%/10% split in test. Moreover, a single feature present in the training data is capable of correctly sorting all samples, which is not biologically feasible as the phenotype is an association of expression patterns, genetic profiles, and environmental factors; thus, it is not regulated by a single variable (Burga & Lehner, 2012; Grishkevich & Yanai, 2013; Jakutis & Stainier, 2021; Wong et al., 2021). However, the same feature is not relevant in the test set, which, in turn, is not linearly separable. Another issue was identified in the data from Bhattacharjee et al. (2001). This adenocarcinoma dataset from 2001 appears as the 11th most used in our ranking. Using outlier analysis and the supplemental information of the original publication, Mramor et al. (2007) identified seven mislabeled samples. These shifts in data distribution and the presence of outliers are known to impact the quality of feature selectors and classifiers Dorn et al. (2021).

Other issues are less evident. The most used dataset found in our review was the leukemia microarray experiment by Golub et al. (1999) from 1999. Containing the expression of 6817 genes and only 38 samples split between two classes, this dataset can be considered small even when compared to data from the early 2000s. It is also highly heterogeneous. It includes samples from peripheral blood rather than just bone marrow, as well as from child patients and from laboratories that used different sample preparation protocols. More alarmingly, using a method called "neighborhood analysis," which is based on the expression levels of individual genes being uniformly high or low, the authors reported 100% accuracy on the class prediction. These results were highly insensitive to the particular selection of genes, with predictors using different inputs ranging from 10 to 200 genes, all achieving the same 100% accuracy. There are no doubts about the relevance and pioneering of Golub et al. (1999) results. Nevertheless, it is striking that a small dataset from 1999 with classes that can be perfectly classified using several assortments of genes is still the most popular benchmark for new feature selection studies (40% of the reviewed publications).

Another possible cause for concern in some datasets is that they were prefiltered. The leukemia and the SRBCT datasets from Alon et al. (1999) and Khan et al. (2001) placed second and fifth in our rank of most used datasets. Both used microarray experiments with 6567 genes. However, the authors reduced the number of genes in their data to 2000 and 2308, respectively, by filtering for a minimal expression level. In this sense, the versions of these two datasets that ended up being employed by other researchers were their filtered variations, not the original raw data (Bolón-Canedo et al., 2014). Although not a problem, using prefiltered data to evaluate feature selectors can bias the results. By using data that had several features discarded, the actual dimension of the task is reduced (approximately three times in these examples), and it becomes impossible to evaluate how the selection algorithms would deal with irrelevant or redundant features present in the original data. For benchmarking purposes, the authors may be inadvertently mixing the results of a third-party filter algorithm with their own.

A final problem with these older datasets is the accessibility of the complete list of gene names, accession numbers, and original expression matrices. For example, Pomeroy et al. (2002) employed a microarray platform with only 6817 probes, which predates the end of the HGP and is three times smaller than the frequently employed Affymetrix U133 GeneChip. The same platform was used by Shipp et al. (2002). Likewise, the links to the original gene expression results of Pomeroy et al. (2002) and the detailed gene expression analysis protocol of Armstrong et al. (2002) were stored in private websites that are no longer functional—a common issue with datasets published before GEO. Finally, another issue of the employed older datasets (e.g., Van't Veer et al., 2002) is the information of cRNA sequences for microarrays, which were extracted before the end of the HGP, thus, containing several biases by current standards.

A necessary mindset change must be reached to ensure that newer feature selection studies do not perpetuate inaccurate and outdated data usage. Older datasets suffer from numerous inadequacies for today's standards. From discontinued platforms and erroneous probes to unknown sources and obscure processing steps, these datasets were

continuously employed in divergent biological questions, generating doubtful biological data. Likewise, the inherent heterogeneous nature of cancer, which is rarely accounted for in feature selection or machine learning studies, must also be evaluated when combining different datasets. Finally, researchers should always know the peculiarities of RNA-seq and microarray expression matrices.

5 | AVAILABLE DATABASES: A PANDORA'S BOX

Another challenge in using gene expression data in bioinformatics and computational biology is finding trustworthy sources that make them available. Thus, one of the core questions in feature selection applied to gene expression data is where the researchers find datasets to perform experiments. Bolón-Canedo et al. (2014) listed what were considered the nine most famous microarray data repositories in 2014. This list is reproduced in Table 3, with their current availability status. Of the nine databases, only three remain fully accessible through the URLs listed in 2014. This situation showcases some of the permanent challenges of choosing the sources of datasets in feature selection research. Databases with little supervision carry the risk of lack of maintenance or even deletion. This also hinders the analysis of past publications and comparison with older results because critical data or metadata may be missing.

Over the years, numerous large machine learning databases became available to discover innovative techniques to perform learning tasks. Some well-known general platforms and dataset resources are the UCI Machine Learning repository (https://archive.ics.uci.edu/ml/index.php) (Dua & Graff, 2017), Kaggle (https://www.kaggle.com/datasets), and OpenML (https://openml.org/). These databases contain hundreds to thousands of datasets from various domains. However, given that these platforms' main goal is the ease of distribution and access of datasets focusing on machine learning experiments and benchmarking and do not necessarily focus on specific domains, they may amplify certain biases by promoting already popular datasets. A search for "gene expression" on Kaggle on July 18, 2022 returned as first result the "Gene expression dataset (Golub et al.)" (https://www.kaggle.com/datasets/crawford/gene-expression). This dataset is a copy of the leukemia dataset from Golub et al. (1999), uploaded to Kaggle in 2017 (Figure 3 in Supplementary Material 1). Since then, it has received over 100,000 views and was downloaded over 12,000 times. The dataset webpage describes it as a good test for classification algorithms without referencing the issues discussed in Sections 3 and 4. Another copy of this dataset appears as one of the first results from OpenML when searching for the keywords "cancer" or "gene expression" (search conducted on July 18, 2022), with even fewer details (https://openml.org/search?type=data&id=1104). The UCI Machine Learning repository contains 147 datasets related to life sciences (search conducted on July 19, 2022), from which one is a human cancer RNA-seq gene expression experiment (https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq).

On the other end of the spectrum, some databases contain specific biological data. The NCBI GEO DataSets (https://www.ncbi.nlm.nih.gov/gds) and the EMBL-EBI ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) are the de facto repositories of gene expression experiment results, in which researchers deposit new datasets. The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov/) is also a well-known database for cancer RNA-seq data, but TCGA does not allow users to store new data continuously. TCGA was an NCI-funded large-scale project and is distinct from the other resources. Nevertheless, all those are biology-first databases with organization, jargon, and file formats that may drive off users only interested in finding datasets to test feature selection and machine learning algorithms.

To bridge the gap between biology and computer science, some databases provide curated gene expression datasets for use in machine learning research, such as DataMicroArray (https://github.com/ramhiser/datamicroarray), BioLab (Mramor et al., 2007) (https://file.biolab.si/biolab/supp/bi-cancer/projections/), Princeton University Gene Expression Project (http://genomics-pubs.princeton.edu/oncology/), inSilicoDB (Taminau et al., 2011) (https://www.bioconductor.org/packages//2.10/bioc/html/inSilicoDb.html), PSO-EMT datasets (Chen et al., 2020) (https://ckzixf.github.io/dataset.html), the CuMiDa (Feltes et al., 2019) (https://sbcb.inf.ufrgs.br/cumida) with microarray data, and the BARRA:CuRDa (Feltes et al., 2021) (https://sbcb.inf.ufrgs.br/barracurda) with curated RNA-seq data. Table 4 compares these databases. Most of them are created and organized by specific research groups to share data used in their experiments. With a few exceptions, they are not updated with newer datasets, nor do they discuss the biological relevance or particularities of the data. However, they are a valuable source for feature selection experiments—the files are freely and openly shared in ready-to-use formats compatible with the most popular data science libraries and software. Moreover, they come with the legitimacy of being used in past feature selection or machine learning publications. With the need to provide enough data to train machine learning models, the few database resources became rapidly famous and broadly used by

TABLE 3 The most famous public microarray data repositories according to the review by Bolón-Canedo et al. (2014).

Repositories	URL	Status
ArrayExpress European Bioinformatics Institute	http://www.ebi.ac.uk/arrayexpress/	✓
Cancer Program Data Sets Broad Institute	http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi	✖
Dataset Repository Bioinformatics Research Group of Universidad Pablo de Olavide	http://www.upo.es/eps/bigs/datasets.html	✖
Feature Selection Datasets Arizona State University	http://featureselection.asu.edu/datasets.php	✖
Gene Expression Model Selector Vanderbilt University	http://www.gems-system.org	✖
Gene Expression Omnibus National Institutes of Health	http://www.ncbi.nlm.nih.gov/geo/	✓
Gene Expression Project Princeton University	http://genomics-pubs.princeton.edu/oncology/	✓
Kent Ridge Bio-Medical Dataset Repository Agency for Science, Technology and Research	http://datam.i2r.a-star.edu.sg/datasets/krbd	✖
Stanford Microarray Database Stanford University	http://smd.stanford.edu/	✖

Note: The sources and the provided URLs are listed as informed by Bolón-Canedo et al. (2014), which accessed all repositories in January 2014. In less than a decade, most URLs are not working anymore (marked with 55). The current status of the URLs was checked on October 14, 2023.

TABLE 4 Comparison between some prominent gene expression databases from which feature selection researchers are likely to download datasets for experiments.

Databases	Curated	Source	Quality control[a]	Up to date[b]	File formats[c]
`ARCH4`	No	Normalized; gene and transcript level counts	No	Yes	`.h5`
`BARRA:CuRDa`	Yes	Normalized	Yes	Yes	`.csv; .tab; .gct; .cls`
`BioLab`	No	Author's	No	No	`.tab`
`CuMiDa`	Yes	Normalized	Yes	Yes	`.csv; .tab; .gct; .cls; .arff`
`Datamicroarray`	No	Author's	No	No	`.r; .RData`
`Gene Expression Project`	No	Author's	No	No	`.tab; .xls`
`InSilicodb`	Yes	Varies	NS	Yes	`.r`
`PSO-EMT datasets`	No	Varies	No	No	`.mat`
`Recount3`	Yes	Gene and transcript level counts	No	Yes	`.txt; .bw; .mtx; .RData`
`Refine.bio`	No	Normalized; gene and transcript level counts	No	Yes	`.sf; .json; .tsv`
`RNASeq-er`	No	Normalized; gene and transcript level counts	No	Yes	`.cram; .bw; .bedGraph`

Note: The databases are listed in alphabetical order. This table was updated and adapted from Feltes et al. (2019). In this case, since *inSilicodb* offers datasets curated by the community, the condition they were built depends on the user.

Abbreviations: BARRA:CuRDa, Benchmarking of ARtificial intelligence Research: Curated RNA-seq Database. CuMiDa, Curated Microarray Database; NS, not specified.

[a]Referring to low-quality sample exclusion.

[b]We are considering databases that offer datasets from the last 5 years or if most of the datasets are at least from the last 10 years.

[c]Some databases, such as *inSilicodb* and *datamicroarray*, which are R packages, can be exported in different formats due to R flexibility. In this case, we only list the default entries they offer or their regular file format. *inSilicodb* does not possess a file format since the information is imported directly into R.

the machine learning community, often without the proper quality control. As a consequence, some of them end up perpetuating the sharing of the same datasets listed in Table 2.

Different projects have worked to make available processing raw data from publicly RNA-seq or microarray datasets, allowing cross-validation and supporting post hoc analysis from multiple organisms. Among these projects are recount3 (Wilks et al., 2021), ARCH4 (Lachmann et al., 2018), refine.bio (Greene et al., 2023), and RNASeq-er (Petryszak et al., 2017), which provide the gene counting summarization at gene and transcript levels. However, while the data available in the repositories listed above have the great advantage of the number of samples, they are not focused on machine learning approaches, thus, their matrices still need prior steps to ensure sample consistency and proper prior normalization.

The CuMiDa (Feltes et al., 2019; Grisci et al., 2019), which is strictly focused on cancer research, was created with the challenges related to analyzing gene expression data with feature selection-based algorithms in mind. CuMiDa stands out from other known databases by trying to solve the issues mentioned in the previous sections, which are: (i) the age of the datasets; (ii) proper handling (preprocessing); (iii) dealing with biological reality; and (iv) benchmarking. Differently from the datasets listed on all other similar databases, the ones in CuMiDa are derived from the manual curation of all cancer-associated microarray datasets of the entire GEO database. The curation used rigorous filtering criteria to exclude technical artifacts, low-quality samples, and faulty probes. From more than 30,000 manually curated datasets, only 78 fitted all criteria and filtering, showing how difficult it is to find quality datasets that would optimally answer a biological question. However, we point out that CuMiDa relies only on datasets with samples from *H. sapiens* and from a specific background (human cancers), and, thus, it does not solve the issue of lack of diversity in data by itself. A similar database named BARRA:CuRDa (Feltes et al., 2021), a sister database of CuMiDa, focused on RNA-seq. BARRA:CuRDa was meant to tackle another overlooked issue: RNA-seq preprocessing for ML. As described before, RNA-seq matrices are vastly different from microarrays and should be handled in their particular way. The 17 datasets in BARRA:CuRDa were derived from similar filtering but analyzed using state-of-the-art RNA-seq preprocessing measures for alignment, trimming, and read quantification and quality.

6 | CONCLUSION

There is a tendency of feature selection works applied to gene expression data to follow a mindset that could be detrimental to the field. From 1284 works from the last 11 years, we observed that: (i) 32.3% never cited the original work, referencing only intermediate articles, showing that there is a tendency to replicate data without proper knowledge of the background thematic; (ii) 7.7% were broken links, making the datasets they used unreachable; (iii) 5.1% never cited any source; (v) only 4.6% were generated by the authors. Older datasets that are outdated and, biologically speaking, should no longer be used to convey biological information are also a constant concern. In this sense, 40.26% of the reviewed publications from the last 11 years used the data from Golub et al. (1999) that, despite pioneering the field, is outdated and from a platform discontinued for the last 15–20 years. Likewise, the works of Alon et al. (1999) and Singh et al. (2002) encompassed 32.13% and 22.98% of the employed datasets. Unfortunately, we noted that using outdated data are still a trend. Although understandably, multiple Computer Science-oriented works are more focused on the practical aspects of the data. Most claim that the algorithm can be used to convey biological information, which is inaccurate and misleading due to the multitude of technical and biological issues previously discussed. More importantly, these issues are not due to the lack of available new data but to a harmful mindset that keeps being perpetuated.

Hence, creating clear and objective guidelines in the field is paramount to ensure that feature selection algorithms can be safely applied to biological datasets and generate biologically meaningful data. Some mandatory recommendations are: (i) always cite the original data source and do not use only indirect citations. (ii) the main characteristics of the datasets, such as the number of samples, features, classes, and so on, should always be mentioned. If several datasets are used, list them in an organized way, such as a table; (iii) add working hyperlinks to the data source. If the data are stored in private databases, check if they are not also available in public repositories. If they are, cite the public repository as well to assure that it can be reached from multiple sources; (iv) use several datasets with distinct properties to validate new algorithms; (v) explore the use of datasets from model organisms other than *H. sapiens* and conditions other than cancer to avoid biases in the evaluation of new algorithms intended for general analyses; (vi) prevent the use of author's pre-filtered datasets, as they may bias the results of new experiments and algorithms; (vii) avoid using only older datasets. They can be used for comparison with previous works, but current datasets should always be employed to validate experiments due to the increasing biological data that is constantly changing; (viii) explore RNA-

seq data, keeping in mind that microarray expression matrices are different from RNA-seq read count matrices; (ix) when designing a new database of gene expression for feature selection or machine learning applications, follow the guidelines proposed by Peters et al. (2018), Wilkinson et al. (2016), Walsh et al. (2021), and Hutchinson et al. (2021); (x) reviewers and editors should enforce the items above in the publications. Another relevant trend is to use interpretable machine learning and visualization methods (Artur & Minghim, 2019; Dennig et al., 2019; Grisci et al., 2021; May et al., 2011) to improve the understanding and replicability of the results of feature selection experiments with high-dimensional data. Even though this work focused on feature selection, given the scope of the reviewed publications, many of these recommendations also apply to machine learning and bioinformatics research applied to gene expression data.

AUTHOR CONTRIBUTIONS

Bruno I. Grisci: Conceptualization (equal); formal analysis (equal); supervision (equal); writing – original draft (equal). **Bruno César Feltes:** Conceptualization (equal); formal analysis (equal); visualization (equal); writing – original draft (equal). **Joice de Faria Poloni:** Conceptualization (equal); formal analysis (equal); visualization (equal); writing – original draft (equal). **Pedro H. Narloch:** Formal analysis (equal); writing – original draft (equal). **Márcio Dorn:** Funding acquisition (equal); supervision (equal); writing – original draft (equal).

FUNDING INFORMATION

This work was supported by grants from the Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul—FAPERGS (19/2551-0001906-8 and 17/2551-0000520-1), Conselho Nacional de Desenvolvimento Científico e Tecnológico—CNPq (440279/2022-4; 408154/2022-5, 314082/2021-2, 465450/2014-8, and 151591/2022-9), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—CAPES/STICAMSUD (88881.522073/2020-01) and DAAD/CAPES PROBRAL (88881.198766/2018-01), and the Emerging Leaders in the Americas Program Scholarship with the support of the Government of Canada. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brazil (CAPES)—Finance Code 001.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Bruno I. Grisci https://orcid.org/0000-0003-4083-5881
Márcio Dorn https://orcid.org/0000-0001-8534-3480

RELATED WIREs ARTICLES

Identification of significant features in DNA microarray data

REFERENCES

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., ... Staudt, L. M. (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, *403*, 503–511.

Allison, D. B., Cui, X., Page, G. P., & Sabripour, M. (2006). Microarray data analysis: From disarray to consolidation and consensus. *Nature Reviews Genetics*, *7*, 55–65.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, *96*, 6745–6750.

Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. A. (2016). Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *Institute of Electrical and Electronics Engineers/Association for Computing Machinery Transactions on Computational Biology and Bioinformatics*, *13*, 971–989.

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R., & Korsmeyer, S. J. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, *30*, 41–47.

Artur, E., & Minghim, R. (2019). A novel visual approach for enhanced attribute analysis and selection. *Computers & Graphics*, *84*, 160–172.

Beer, D. G., Kardia, S. L., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M., Iannettoni, M. D., Orringer, M. B., & Hanash, S. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, *8*, 816–824.

Beker, W., Roszak, R., Wolos, A., Angello, N. H., Rathore, V., Burke, M. D., & Grzybowski, B. A. (2022). Machine learning may sometimes simply capture literature popularity trends: A case study of heterocyclic Suzuki–Miyaura coupling. *Journal of the American Chemical Society*, *144*, 4819–4827.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J., & Meyerson, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 13790–13795.

Blalock, E. M. (2003). *A beginner's guide to microarrays*. Springer Science & Business Media.

Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Information Sciences*, *282*, 111–135.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144–152). Association for Computing Machinery.

Boulesteix, A.-L., Strobl, C., Augustin, T., & Daumer, M. (2008). Evaluating microarray-based classifiers: An overview. *Cancer Informatics*, *6*, 77–97.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge.

Burga, A., & Lehner, B. (2012). Beyond genotype to phenotype: Why the phenotype of an individual cannot always be predicted from their genome sequence and the environment that they experience. *The Federation of European Biochemical Societies Journal*, *279*, 3765–3775.

Carvalho, B. S., & Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, *26*, 2363–2367.

Chain, B. (2021). agilp: Agilent expression array processing package. R package version 3.26.0.

Chen, K., Xue, B., Zhang, M., & Zhou, F. (2020). An evolutionary multitasking-based feature selection method for high-dimensional classification. *Institute of Electrical and Electronics Engineers Transactions on Cybernetics*, *52*, 7172–7186.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., & Foa, R. (2004). Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, *103*, 2771–2778.

Chowdary, D., Lathrop, J., Skelton, J., Curtin, K., Briggs, T., Zhang, Y., Yu, J., Wang, Y., & Mazumder, A. (2006). Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. *The Journal of Molecular Diagnostics*, *8*, 31–39.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*, 1–19.

Corchete, L. A., Rojas, E. A., Alonso-López, D., De Las Rivas, J., Gutiérrez, N. C., & Burguillo, F. J. (2020). Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Scientific Reports*, *10*, 1–15.

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., Bunney, W. E., Myers, R. M., Speed, T. P., Akil, H., Watson, S. J., & Meng, F. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research*, *33*, e175.

Dennig, F. L., Polk, T., Lin, Z., Schreck, T., Pfister, H., & Behrisch, M. (2019). FDive: Learning relevance models using pattern-based similarity measures. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 69–80). IEEE.

Dorn, M., Grisci, B. I., Narloch, P. H., Feltes, B. C., Avila, E., Kahmann, A., & Alho, C. S. (2021). Comparison of machine learning techniques to handle imbalanced COVID-19 CBC datasets. *PeerJ Computer Science*, *7*, e670.

Du, P., Kibbe, W., & Lin, S. (2008). lumi: A pipeline for processing illumina microarray. *Bioinformatics*, *24*, 1547–1548.

Dua, D., & Graff, C. (2017). *UCI machine learning repository*. http://archive.ics.uci.edu/ml

Dunning, M., Smith, M., Ritchie, M., & Tavaré, S. (2007). beadarray: R classes and methods for illumina bead-based data. *Bioinformatics*, *23*, 2183–2184.

Dyrskjøt, L., Thykjaer, T., Kruhøffer, M., Jensen, J. L., Marcussen, N., Hamilton-Dutoit, S., Wolf, H., & Ørntoft, T. F. (2003). Identifying distinct classes of bladder carcinoma using microarrays. *Nature Genetics*, *33*, 90–96.

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*, 207–210.

Epstein, C., & Butow, R. (2000). Microarray technology-enhanced versatility, persistent challenge. *Current Opinion in Biotechnology*, *11*, 36–41.

Feltes, B. C., Chandelier, E. B., Grisci, B. I., & Dorn, M. (2019). CuMiDa: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. *Journal of Computational Biology*, *26*, 376–386.

Feltes, B. C., Grisci, B. I., de Faria Poloni, J., & Dorn, M. (2018). Perspectives and applications of machine learning for evolutionary developmental biology. *Molecular Omics*, *14*, 289–306.

Feltes, B. C., Poloni, J. D. F., & Dorn, M. (2021). Benchmarking and testing machine learning approaches with BARRA:CuRDa, a curated RNA-seq database for cancer research. *Journal of Computational Biology*, *28*, 931–944.

Gautier, L., Cope, L., Bolstad, B., & Irizarry, R. (2004). affy—Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, *20*, 307–315.

Gautier, L., Møller, M., Friis-Hansen, L., & Knudsen, S. (2004). Alternative mapping of probes to genes for Affymetrix chips. *BMC Bioinformatics*, *5*, 1–7.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, *286*, 531–537.

Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J., & Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, *62*, 4963–4967.

Greene, C. S., Hu, D., Jones, R. W., Liu, S., Mejia, D. S., Patro, R., Piccolo, S. R., Romero, A. R., Sarkar, H., Savonen, C. L. et al. (2023) *refine. bio: A resource of uniformly processed publicly available gene expression datasets*. https://www.refine.bio

Grisci, B. I., Feltes, B. C., & Dorn, M. (2018). Microarray classification and gene selection with fs-neat. In *2018 IEEE Congress on Evolutionary Computation (CEC)* (pp. 1–8). IEEE.

Grisci, B. I., Feltes, B. C., & Dorn, M. (2019). Neuroevolution as a tool for microarray gene expression pattern identification in cancer research. *Journal of Biomedical Informatics*, *89*, 122–133.

Grisci, B. I., Krause, M. J., & Dorn, M. (2021). Relevance aggregation for neural networks interpretability and knowledge discovery on tabular data. *Information Sciences*, *559*, 111–129.

Grishkevich, V., & Yanai, I. (2013). The genomic determinants of genotype × environment interactions in gene expression. *Trends in Genetics*, *29*, 479–487.

Harbig, J., Sprinkle, R., & Enkemann, S. A. (2005). A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Research*, *33*, e31.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362.

Haslinger, C., Schweifer, N., Stilgenbauer, S., Dohner, H., Lichter, P., Kraut, N., Stratowa, C., & Abseher, R. (2004). Microarray gene expression profiling of B-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and VH mutation status. *Journal of Clinical Oncology*, *22*, 3937–3949.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., ... Trent, J. (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, *344*, 539–548.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278–282). IEEE.

Holmes, G., Donkin, A., & Witten, I. H. (1994). Weka: A machine learning workbench. In *Proceedings of ANZIIS'94-Australian New Zealnd Intelligent Information Systems Conference* (pp. 357–361). IEEE.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., ... Morgan, M. (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nature Methods*, *12*, 115–121.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., & Mitchell, M. (2021). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 560–575). Association for Computing Machinery.

Jakutis, G., & Stainier, D. Y. (2021). Genotype–phenotype relationships in the context of transcriptional adaptation and genetic robustness. *Annual Review of Genetics*, *55*, 71–91.

Kandaswamy, K. K., Pugalenthi, G., Hazrati, M. K., Kalies, K.-U., & Martinetz, T. (2011). BLProt: Prediction of bioluminescent proteins based on support vector machine and relieff feature selection. *BMC Bioinformatics*, *12*, 1–7.

Kauffmann, A., Gentleman, R., & Huber, W. (2008). arrayQualityMetrics—A bioconductor package for quality assessment of microarray data. *Bioinformatics*, *25*, 415–416.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., & Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, *7*, 673–679.

Koch, B., Denton, E., Hanna, A., & Foster, J. (2021). Reduced, reused and recycled: The life of a dataset in machine learning research. In *NeurIPS dataset and benchmark track*. NeurIPS.

Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., Wang, L., Silverstein, M. C., & Ma'ayan, A. (2018). Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications*, *9*, 1366.

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., & Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. *Institute of Electrical and Electronics Engineers/Association for Computing Machinery Transactions on Computational Biology and Bioinformatics*, *9*, 1106–1119.

Leban, G., Zupan, B., Vidmar, G., & Bratko, I. (2006). VizRank: Data visualization guided by machine learning. *Data Mining and Knowledge Discovery*, *13*, 119–136.

Leung, Y. F., & Cavalieri, D. (2003). Fundamentals of cDNA microarray data analysis. *Trends in Genetics*, *19*, 649–659.

Li, C., & Xu, J. (2019). Feature selection with the Fisher score followed by the maximal clique centrality algorithm can accurately identify the hub genes of hepatocellular carcinoma. *Scientific Reports*, *9*, 1–11.

Liang, W., Tadesse, G. A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., & Zou, J. (2022). Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*, *4*, 669–677.

Liu, H., Bebu, I., & Li, X. (2010). Microarray probes and probe sets. *Frontiers in Bioscience (Elite Edition)*, *2*, 325–338.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*, 1–21.

May, T., Bannach, A., Davey, J., Ruppert, T., & Kohlhammer, J. (2011). Guiding feature subset selection with an interactive visualization. In *In 2011 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 111–120). IEEE.

Mramor, M., Leban, G., Demšar, J., & Zupan, B. (2007). Visualization-based cancer microarray data classification analysis. *Bioinformatics*, *23*, 2147–2154.

Nature. (2022). The rise and fall (and rise) of datasets. *Nature Machine Intelligence*, *4*, 1–2.

Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv preprint arXiv:2103.14749.

Notterman, D. A., Alon, U., Sierk, A. J., & Levine, A. J. (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Research*, *61*, 3124–3130.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*, 44–53.

Nutt, C. L., Mani, D., Betensky, R. A., Tamayo, P., Cairncross, J. G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M. E., Batchelor, T. T., Black, P. M., von Deimling, A., Pomeroy, S. L., Golub, T. R., & Louis, D. N. (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research*, *63*, 1602–1607.

Osama, S., Shaban, H., & Ali, A. A. (2022). Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review. *Expert Systems with Applications*, *118946*, 1–25.

Otto, E., Culakova, E., Meng, S., Zhang, Z., Xu, H., Mohile, S., & Flannery, M. A. (2022). Overview of Sankey flow diagrams: Focusing on symptom trajectories in older adults with advanced cancer. *Journal of Geriatric Oncology*, *13*, 742–746.

Pedregosa, F., Varoquaux, G., Duchesnay, E., et al. (2011). scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Institute of Electrical and Electronics Engineers Transactions on Pattern Analysis and Machine Intelligence*, *27*, 1226–1238.

Peters, B., Brenner, S., Wang, E., et al. (2018). Putting benchmarks in their rightful place: The heart of computational biology. *PLoS Computational Biology*, *14*, e1006494.

Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., Moore, R., McClanahan, T. K., Sadekova, S., & Klappenbach, J. A. (2017). Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology*, *35*, 936–939.

Petryszak, R., Fonseca, N. A., Füllgrabe, A., Huerta, L., Keays, M., Tang, Y. A., & Brazma, A. (2017). The RNASeq-er API-a gateway to systematically updated analysis of public RNA-seq data. *Bioinformatics*, *33*, 2218–2220.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y., Goumnerova, L. C., Black, P. M., Lau, C., Allen, J. C., Zagzag, D., Olson, J. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., … Golub, T. R. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, *415*, 436–442.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Gerald, W., Loda, M., Lander, E. S., & Golub, T. R. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 15149–15154.

Ranjan, B., Sun, W., Park, J., Mishra, K., Schmidt, F., Xie, R., Alipour, F., Singhal, V., Joanito, I., Honardoost, M. A., Yong, J. M. Y., Koh, E. T., Leong, K. P., Rayan, N. A., Lim, M. G. L., & Prabhakar, S. (2021). DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data. *Nature Communications*, *12*, 1–12.

Ritchie, M., Phipson, B., Wu, D., Hu, Y., Law, C., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*, e47.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., AIX-COVNET, Ruggiero, A., Korhonen, A., Jefferson, E., Ako, E., Langs, G., … Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, *3*, 199–217.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edger: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*, 139–140.

Robnik-Šikonja, M., & Kononenko, I. (1997). An adaptation of relief for attribute estimation in regression. In *Machine learning: Proceedings of the fourteenth international conference (ICML'97)* (Vol. 5, pp. 296–304). Association for Computing Machinery.

Roman, D., Saxena, S., Robu, V., Pecht, M., & Flynn, D. (2021). Machine learning pipeline for battery state-of-health estimation. *Nature Machine Intelligence*, *3*, 447–456.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., … Lymphoma/Leukemia Molecular Profiling Project. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine*, *346*, 1937–1947.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., & Brown, P. O. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, *24*, 227–235.

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*, 2507–2517.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). Association for Computing Machinery.

Santosa, F., & Symes, W. W. (1986). Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, *7*, 1307–1330.

Shi, H., Zhou, Y., Jia, E., Pan, M., Bai, Y., & Ge, Q. (2021). Bias in RNA-seq library preparation: Current challenges and solutions. *BioMed Research International*, *2021*, 1–11.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C., & Golub, T. R. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, *8*, 68–74.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., & Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, *1*, 203–209.

Smith, L. M., Baggerly, A. K., Bengtsson, H., Ritchie, E. M., & Hansen, D. K. (2013). illuminaio: An open source idat parsing tool for illumina microarrays. *F1000Research*, *264*, 2.

Staunton, J. E., Slonim, D. K., Coller, H. A., Tamayo, P., Angelo, M. J., Park, J., Scherf, U., Lee, J. K., Reinhold, W. O., Weinstein, J. N., Mesirov, J. P., Lander, E. S., & Golub, T. R. (2001). Chemosensitivity prediction by transcriptional profiling. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 10787–10792.

Stienstra, R., Saudale, F., Duval, C., Keshtkar, S., Groener, J. E., van Rooijen, N., Staels, B., Kersten, S., & Müller, M. (2010). Kupffer cells promote hepatic steatosis via interleukin-1β-dependent suppression of peroxisome proliferator-activated receptor α activity. *Hepatology*, *51*, 511–522.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., & Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, *14*, 865–868.

Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A., Patapoutian, A., Hampton, G. M., Schultz, P. G., & Hogenesch, J. B. (2002). Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 4465–4470.

Tadist, K., Najah, S., Nikolov, N. S., Mrabti, F., & Zahi, A. (2019). Feature selection methods and genomic big data: A systematic review. *Journal of Big Data*, *6*, 1–24.

Taminau, J., Steenhoff, D., Coletta, A., Meganck, S., Lazar, C., de Schaetzen, V., Duque, R., Molter, C., Bersini, H., Nowé, A., & Weiss Solís, D. Y. (2011). inSilicoDb: An R/Bioconductor package for accessing human Affymetrix expert-curated datasets fromGEO. *Bioinformatics*, *27*, 3204–3205.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*, 267–288.

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., Van Der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., & Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*, 530–536.

Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Pollastri, G., Harrow, J., Psomopoulos, F. E., & Tosatto, S. C. (2021). Dome: Recommendations for supervised machine learning validation in biology. *Nature Methods*, *18*, 1122–1127.

Welsh, J. B., Sapinoso, L. M., Su, A. I., Kern, S. G., Wang-Rodriguez, J., Moskaluk, C. A., Frierson, H. F., Jr., & Hampton, G. M. (2001). Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Research*, *61*, 5974–5978.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Jr., Marks, J. R., & Nevins, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 11462–11467.

Whiteson, S., Stone, P., Stanley, K. O., Miikkulainen, R., & Kohl, N. (2005). Automatic feature selection in neuroevolution. In *Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation* (pp. 1225–1232). ACM.

Whitworth, G. B. (2010). An introduction to microarray data analysis and visualization. *Methods in Enzymology*, *470*, 19–50.

Wigle, D. A., Jurisica, I., Radulovich, N., Pintilie, M., Rossant, J., Liu, N., Lu, C., Woodgett, J., Seiden, I., Johnston, M., Keshavjee, S., Darling, G., Winton, T., Breitkreutz, B. J., Jorgenson, P., Tyers, M., Shepherd, F. A., & Tsao, M. S. (2002). Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Research*, *62*, 3005–3008.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 1–9.

Wilks, C., Zheng, S. C., Chen, F. Y., Charles, R., Solomon, B., Ling, J. P., Imada, E. L., Zhang, D., Joseph, L., Leek, J. T., Jaffe, A. E., Nellore, A., Collado-Torres, L., Hansen, K. D., & Langmead, B. (2021). recount3: Summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biology*, *22*, 1–40.

Wong, A. K., Sealfon, R. S., Theesfeld, C. L., & Troyanskaya, O. G. (2021). Decoding disease: From genomes to networks to phenotypes. *Nature Reviews Genetics*, *22*, 774–790.

Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C. H., Evans, W. E., Naeve, C., ... Downing, J. R. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, *1*, 133–143.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Grisci, B. I., Feltes, B. C., de Faria Poloni, J., Narloch, P. H., & Dorn, M. (2023). The use of gene expression datasets in feature selection research: 20 years of inherent bias? *WIREs Data Mining and Knowledge Discovery*, e1523. https://doi.org/10.1002/widm.1523