ADVANCED REVIEW

# Supplementary Material 1: The use of gene expression datasets in feature selection research: 20 years of inherent bias?

**Bruno I. Grisci[1,2]* | Bruno César Feltes[1,3]* | Joice de Faria Poloni[4,5]* | Pedro H. Narloch[1] | Márcio Dorn[1,5,6]**

[1]Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, RS, 91501-970, Brazil

[2]Faculty of Computer Science, Dalhousie University, Halifax, NS, B3H 4R2, Canada

[3]Institute of Biosciences, Federal University of Rio Grande do Sul, Porto Alegre, RS, 91501-970, Brazil

[4]School of Health and Life Sciences,Pontifical Catholic University of Rio Grande do Sul, RS, 90619-900, Brazil

[5]National Institute of Science and Technology - Forensic Science, Porto Alegre, RS, Brazil

[6]Center for Biotechnology, Federal University of Rio Grande do Sul, Porto Alegre, RS, 91501-970, Brazil

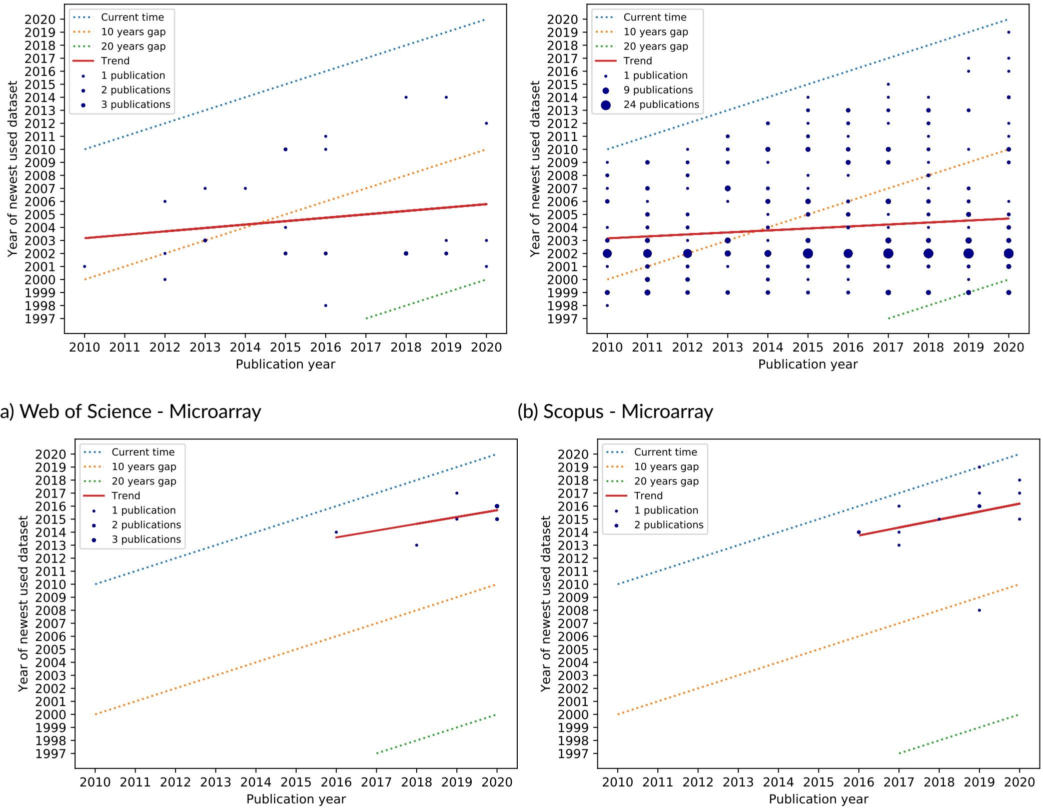
**Correspondence**
Márcio Dorn PhD, Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, RS, 91501-970, Brazil
Email: mdorn@inf.ufrgs.br

**Funding information**
FAPERGS, Grant/Award Number: 19/2551-0001906-8 and 17/2551-0000520-1; CNPq, Grant/Award Number: 314082/2021-2, 465450/2014-8, and 151591/2022-9; CAPES, Grant/Award Number: 88881.522073/2020-01, 88881.198766/2018-01, and Finance Code 001; Government of Canada, Grant/Award Number: ELAP.

Feature selection algorithms are frequently employed in pre-processing machine learning pipelines applied to biological data and identifying relevant features from large-scale datasets. 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selection and gene expression, published between 2010 and 2020, we found that 57% of the publications used at least one outdated dataset, 23% used only outdated data, and 32% did not cite data sources.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.

KEYWORDS
Machine Learning, Feature Selection, Microarray, RNA-seq, Gene Expression

*Equally contributing authors.



(a) Web of Science - Microarray

(b) Scopus - Microarray

(c) Web of Science - RNA-seq

(d) Scopus - RNA-seq

**FIGURE 1** Differences between papers archived in the Web of Science and Scopus databases. The chart details are as in Figure 4 of the main manuscript. (a) year of creation of the most recent microarray dataset used in experiments of papers from Web of Science; (b) year of creation of the most recent microarray dataset used in experiments of papers from Scopus; (c) year of creation of the most recent RNA-seq dataset used in experiments of papers from Web of Science; (d) year of creation of the most recent RNA-seq dataset used in experiments of papers from Scopus.

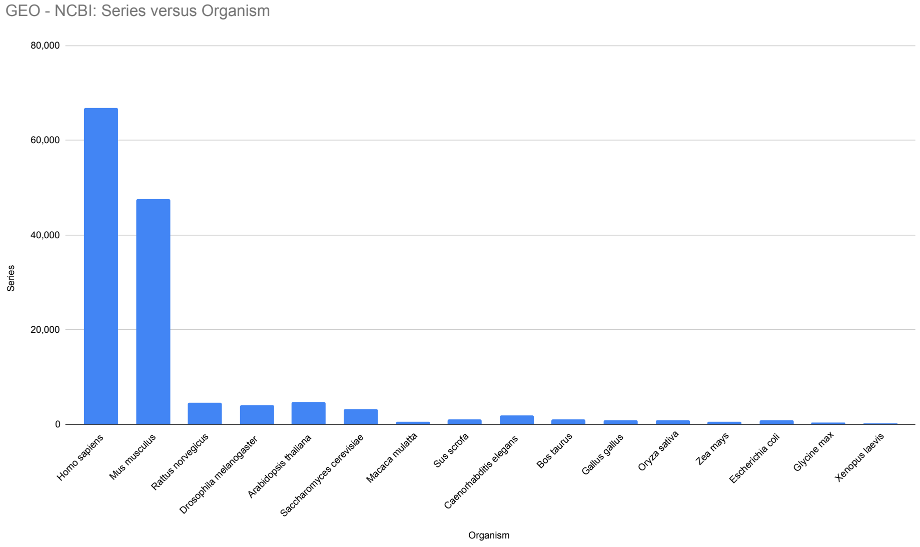
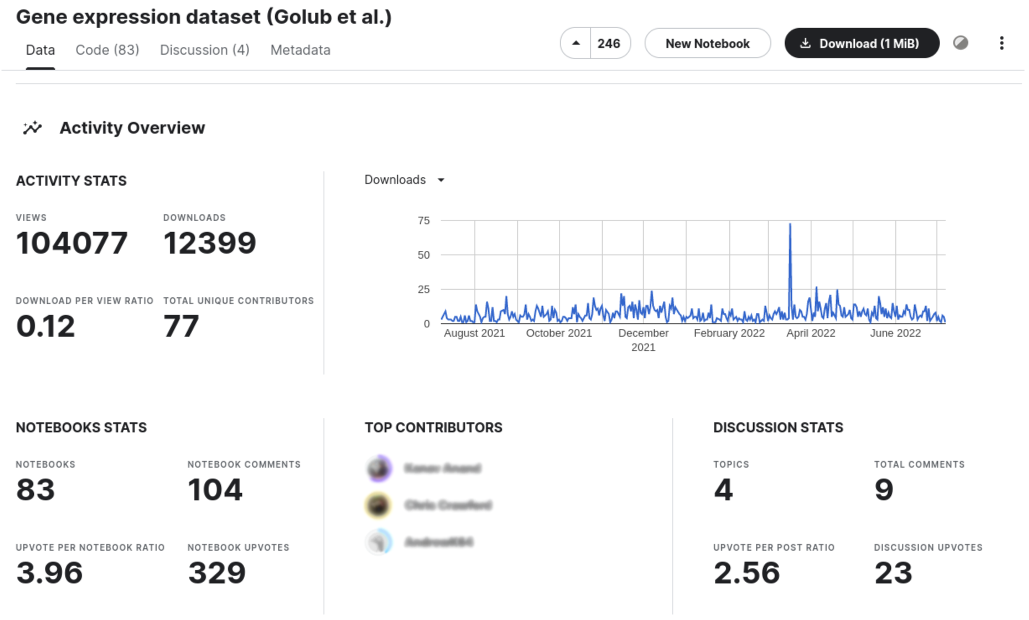


**FIGURE 2** Organism distribution in GEO data series for comparison.


Gene expression dataset (Golub et al.)

Data | Code (83) | Discussion (4) | Metadata

▲ 246 | New Notebook | Download (1 MiB)

Activity Overview

ACTIVITY STATS

VIEWS 104077

DOWNLOADS 12399

DOWNLOAD PER VIEW RATIO 0.12

TOTAL UNIQUE CONTRIBUTORS 77

NOTEBOOKS STATS

NOTEBOOKS 83

NOTEBOOK COMMENTS 104

UPVOTE PER NOTEBOOK RATIO 3.96

NOTEBOOK UPVOTES 329

TOP CONTRIBUTORS

[redacted]

[redacted]

[redacted]

DISCUSSION STATS

TOPICS 4

TOTAL COMMENTS 9

UPVOTE PER POST RATIO 2.56

DISCUSSION UPVOTES 23

**FIGURE 3** Screenshot of the activity overview stats for the "Gene expression dataset (Golub et al.)" published at Kaggle: `https://www.kaggle.com/datasets/crawford/gene-expression`. This reupload of the 1999 leukemia dataset from **?** is one of the first results in the platform after a search for "gene expression," despite the issues discussed in the Section 3 of the main manuscript. The top contributors were blurred for privacy. This screenshot was captured on the July 8th, 2022.