Journal of Digital Imaging (2023) 36:2306–2312
https://doi.org/10.1007/s10278-023-00870-5

Reproducibility of Deep Learning Algorithms Developed for Medical Imaging Analysis: A Systematic Review

Mana Moassefi[1] · Pouria Rouzrokh[1,2] · Gian Marco Conte[1] · Sanaz Vahdati[1] · Tianyuan Fu[3] · Aylin Tahmasebi[4] · Mira Younis[5] · Keyvan Farahani[6] · Amilcare Gentili[7] · Timothy Kline[8] · Felipe C. Kitamura[9] · Yuankai Huo[10] · Shiba Kuanar[1] · Khaled Younis[11] · Bradley J. Erickson[1] · Shahriar Faghani[1]

Received: 16 December 2022 / Revised: 8 June 2023 / Accepted: 9 June 2023 / Published online: 5 July 2023

Abstract

Since 2000, there have been more than 8000 publications on radiology artificial intelligence (AI). AI breakthroughs allow complex tasks to be automated and even performed beyond human capabilities. However, the lack of details on the methods and algorithm code undercuts its scientific value. Many science subfields have recently faced a reproducibility crisis, eroding trust in processes and results, and influencing the rise in retractions of scientific papers. For the same reasons, conducting research in deep learning (DL) also requires reproducibility. Although several valuable manuscript checklists for AI in medical imaging exist, they are not focused specifically on reproducibility. In this study, we conducted a systematic review of recently published papers in the field of DL to evaluate if the description of their methodology could allow the reproducibility of their findings. We focused on the Journal of Digital Imaging (JDI), a specialized journal that publishes papers on AI and medical imaging. We used the keyword "Deep Learning" and collected the articles published between January 2020 and January 2022. We screened all the articles and included the ones which reported the development of a DL tool in medical imaging. We extracted the reported details about the dataset, data handling steps, data splitting, model details, and performance metrics of each included article. We found 148 articles. Eighty were included after screening for articles that reported developing a DL model for medical image analysis. Five studies have made their code publicly available, and 35 studies have utilized publicly available datasets. We provided figures to show the ratio and absolute count of reported items from included studies. According to our cross-sectional study, in JDI publications on DL in medical imaging, authors infrequently report the key elements of their study to make it reproducible.

Keywords Reproducibility · Artificial intelligence · Machine learning · Deep learning · Medical imaging

Abbreviations:

AI	Artificial intelligence
DL	Deep learning
JDI	Journal of Digital Imaging

✉ Mana Moassefi
Moassefi.mana@Mayo.edu

✉ Shahriar Faghani
Faghani.shahriar@mayo.edu

1 Artificial Intelligence Lab, Department of Radiology, Mayo Clinic, Rochester, MN, USA

2 Orthopedic Surgery Artificial Intelligence Laboratory (OSAIL), Department of Orthopedic Surgery, Mayo Clinic, Rochester, MN, USA

3 Department of Radiology, University Hospitals Cleveland, Cleveland, OH, USA

4 Department of Radiology, Thomas Jefferson University, Philadelphia, PA, USA

5 Cleveland Clinic Children's, Cleveland, OH, USA

6 National Cancer Institute, National Institutes of Health, Bethesda, MA, USA

7 Department of Radiology, University of California, San Diego, CA, USA

8 Department of Radiology, Mayo Clinic, Rochester, MN, USA

9 DasaInova, Diagnósticos da América S.A, São Paulo, Brazil

10 Department of Electrical Engineering & Computer Science, Vanderbilt University, Nashville, TN, USA

11 Phillips Research North America, Cambridge, MD, USA

CLAIM Checklist for Artificial Intelligence in Medical Imaging
PRISMA Preferred Reporting Items for Systematic Reviews and Meta-analyses

Introduction

Artificial intelligence (AI) breakthroughs provide enormous potential for automating complex tasks in the field of medical imaging. In recent years, machine learning has been increasingly applied, with more than 8000 radiology AI publications worldwide from 2000 to 2018 [1]. Deep learning (DL) is one of the machine learning methods which uses deeply stacked artificial neurons to perform automatic feature extraction. The application of DL in medical imaging has been extensively investigated. A systematic review found 535 articles on the application of DL in radiology from 2015 to 2019 and showed that this number has exponentially increased over these 4 years [2]. Different journals worldwide publish these articles according to different guidelines [1].

In medical imaging, DL models hold great potential for streamlining clinical workflows [3–5]. It is, however, imperative that the studies that report the development of a DL model should be reproducible to achieve their full potential [6]. DL reproducibility in empirical research refers to an independent team of researchers being able to replicate the results using the same DL methods as the original researchers [7]. The reproducibility crisis in digital medicine has elicited mounting concern within the scientific community [6, 8, 9]. For a study to be truly reproducible, three criteria must be met: technical, statistical, and conceptual reproducibility [10].

Technical reproducibility refers to the ability to reproduce a paper's results precisely as presented in the paper under the same condition [10, 11]. The statistical reproducibility of a study refers to its ability to be replicated under slightly different conditions without statistically significant differences. Different random initializations or random sampling of the data used to train and validate the model, which are related to the internal validity of the model, are a few examples. In this case, even if the results are not identical, they should be statistically equivalent. A model's conceptual reproducibility describes how the desired outcome can be replicated under conditions consistent with its high-level description. Conceptual reproducibility is closely related to external validity [12].

It is essential to increase the reproducibility of AI research to ensure its credibility [7]. The external replication of a study is expected to be more objective. Other researchers have no gain in inflating the performance of a method they have not developed themselves. Their preconceptions and implicit knowledge will differ from those of the first team that reported the research. Therefore, replication of a study is a reasonable step before the clinical application of each DL tool.

Reproducing an experiment requires detailed documentation, which must include relevant information. Whether it is possible to reproduce the experiment's results determines the documentation's relevance and how detailed it must be. Using standard reporting guidelines, researchers can ensure that their final publications contain relevant information [13]. However, to develop a DL algorithm, researchers should carefully design and implement a pipeline of data handling, model development, and model evaluation. Due to the extra elements not conventionally prespecified in traditional reporting guidelines, studies on DL algorithm development have added complexities to how such studies must be reported [14].

There are currently few reporting guidelines for common radiology research studies and their AI-related extensions. For instance, in 2022, the Radiological Society of North America published a Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [15]. CLAIM guideline elements address a wide spectrum of AI applications using medical images. This checklist comprises 42 items, with particular emphasis on data, the reference standard of "ground truth," and the development and methodology of the AI algorithm performance evaluation. In the same year, the American Medical Informatics Association provided a set of guidelines termed the "MI for Medical AI Reporting" (MINIMAR), specific to studies reporting the use of AI solutions in healthcare [16, 14]. Although these checklists cover many aspects of a DL study, they are not focused specifically on reproducibility.

In this systematic review, we focused on papers published in the Journal of Digital Imaging (JDI), a specialized journal that publishes papers on AI and medical imaging. Our team conducted this systematic review to determine how detailed DL studies have been reported in JDI in recent years. We hypothesize that most DL research cannot be reproduced due to inadequate documentation and lack of detailed descriptions in published DL studies.

Methodology

We used the keyword "Deep Learning" to retrieve published studies in JDI between January 2020 and January 2022. Following the title and abstract screening, we included the studies which reported the development of a DL tool in medical imaging. Studies meeting any of the following criteria were excluded: (i) review articles, (ii) articles with only external

validation or usage of other's developed tools, (iii) studies with conventional machine learning techniques, and (iv) not computer vision studies.

Ten reviewers (PR, BK, SV, GC, AT, KY, GF, KF, AG, and YH) performed title, abstract, and full-text screening. Two reviewers reviewed each paper separately, and conflicts were resolved by the third reviewer's opinion (MM or SF). Since making a dataset or code publicly available will address the data and code technical reproducibility aspects. Initially, we verified whether the manuscripts contained any information regarding the release of their data and code. Subsequently, we extracted the reported details about the dataset, data handling steps, data splitting, model details, and performance metrics of each of these articles based on supplementary Table 1. Using these items, we assessed the three reproducibility criteria noted in the Introduction. Conflicts regarding the eligibility of documents during the screening process were resolved by consensus between the reviewers. No bias evaluation was done on the individual search results to maximize the number of included studies. We reported the results according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) extension to scoping reviews [17].

Results

The search on the JDI website resulted in 148 documents. Eighty of them were included after title, abstract, and full-text screening for articles that reported developing a deep learning model for medical image analysis (Fig. 1) (Supplementary Table 2).

Four studies have made their code publicly available, and 35 studies have released their dataset or utilized publicly available datasets. Figures 2 and 3 show the absolute count of studies that reported items in training and external validation from included studies. Figures 4 and 5 display the ratio of studies that reported items in training and external validation from included studies when applicable. For example, 35 studies have used publicly available datasets. For those studies, we have not evaluated reporting the dataset size, time coverage, owner, and inclusion/ exclusion criteria. Another note is that many items were not applicable when performing external validation; for example, reporting the number of epochs, criteria for saving the model, and loss function are not applicable in this case. As a result, studies that underwent external validation have fewer items assessed and documented.

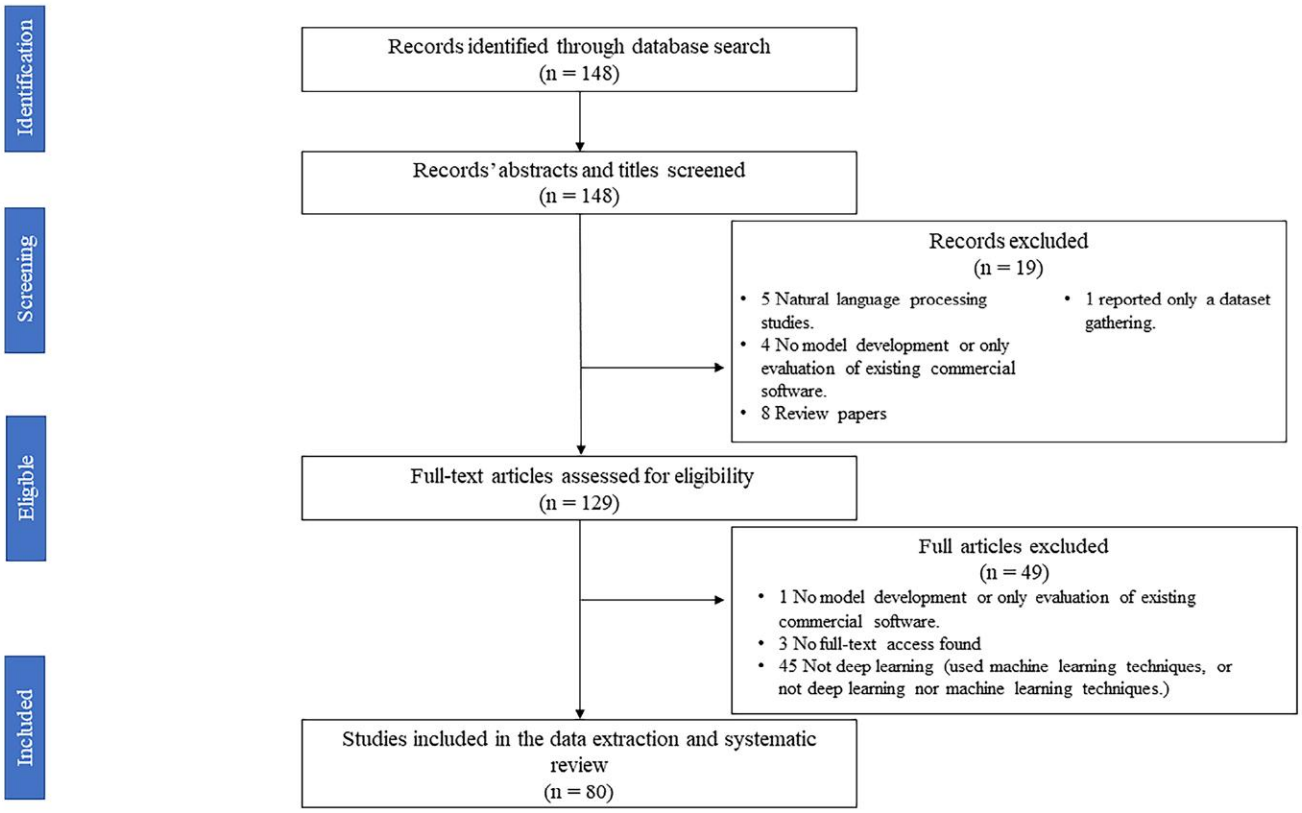

Fig. 1 PRISMA flowchart for the conducted review

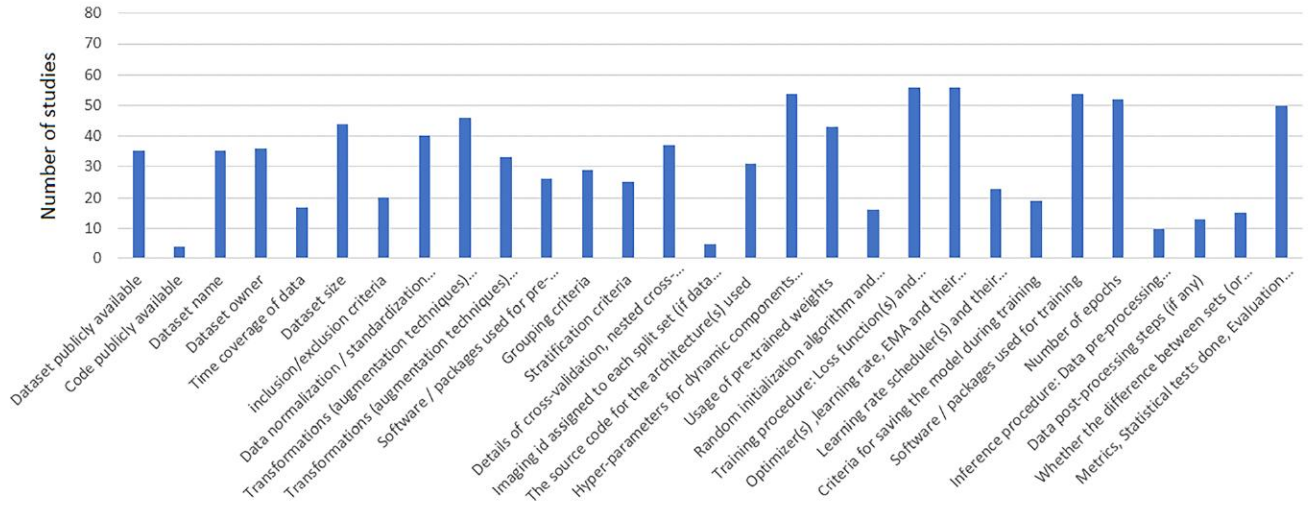

Fig. 2 The illustration of the absolute number of studies that reported particular items on the internal dataset

Discussion

In this systematic review, we screened 148 records and included 80. We limited our search time to January 2020 to January 2022 to extract more detailed items from recently published studies in JDI. We only included the studies in which they described the development of a DL tool and no external validation or studies without the tool development. These articles are the ones that need to be reproduced, and we evaluated the reported details to assess the feasibility of replication.

External validation was done in 5 papers (6.25% of studies) reported; 3 of them used publicly available datasets for performing the external validation tasks. Thirty-five used publicly available datasets and four studies published their code. Based on the extracted items, more than half of the studies reported these items; dataset owner, dataset size, qualitative report of augmentation techniques, details of cross-validation, hyper-parameters for training, training procedure, optimizers, learning rate, software or packages used for training, number of epochs, statistics for performance evaluation.

In studies with internal datasets only, the top five highest reported items were "dataset size," "dataset owner," "Hyper-parameters for dynamic components of the model's architecture (batch norm layers, dropout probability, etc.)," "Training procedure: Loss function(s) and their hyperparameter," and "Optimizer(s), learning rate, exponential moving average (EMA) and their hyperparameter," respectively. The fewest reported items in studies were "dataset names,"

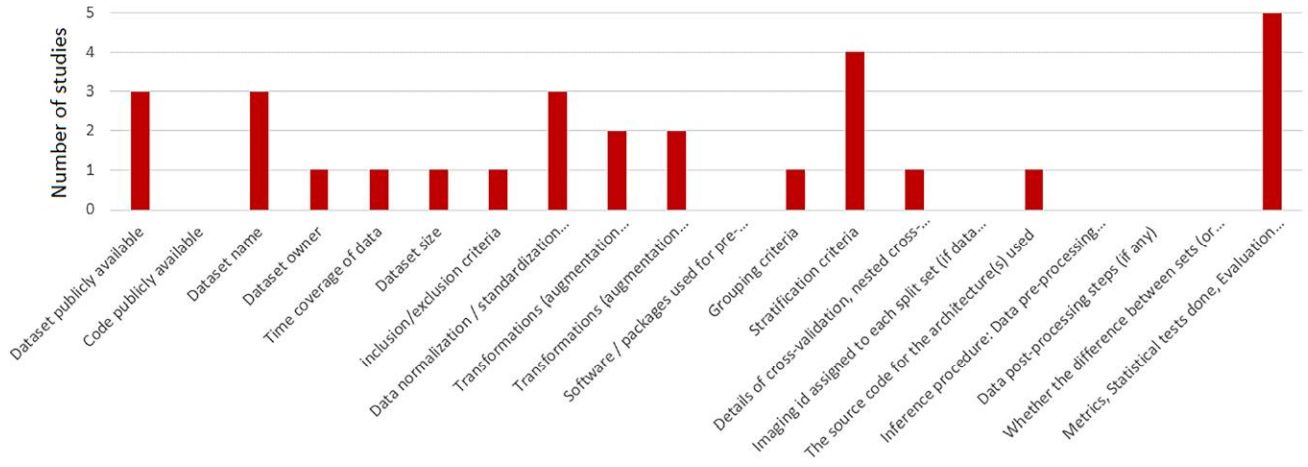

Fig. 3 The illustration of the absolute number of studies that reported particular items on the external dataset

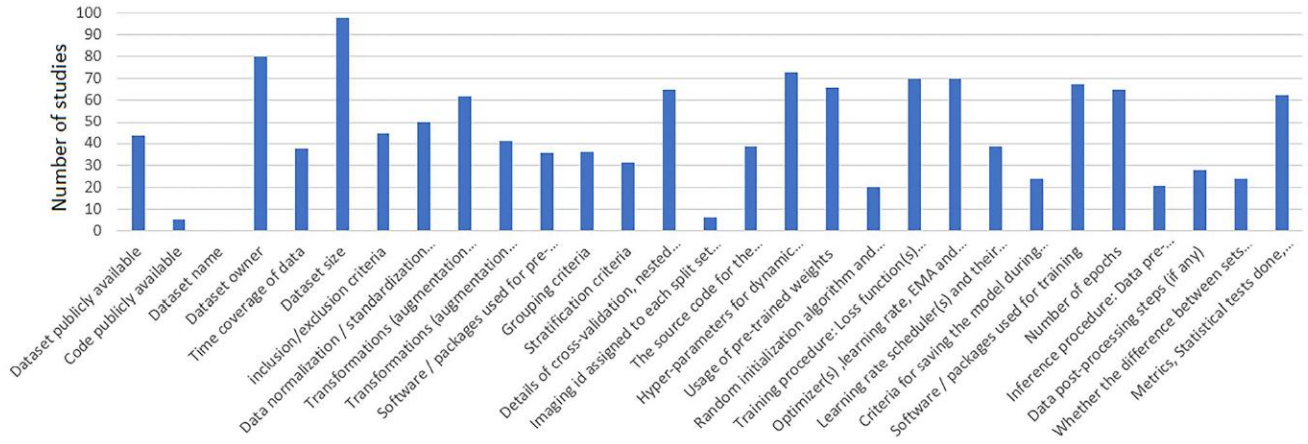

Fig. 4 The illustration of the frequency of the reported items on the internal dataset when applicable

"Imaging id assigned to each split set (if data is public) or seed number used for splitting (if data is private)," "Code publicly available," "Random initialization algorithm and specification and inference procedure: Data pre-processing steps (if different from training)." In the section pertaining to the dataset name and owner, it is worth mentioning that when referring to the dataset owner, we may specify the institution that possesses the dataset. However, it should be noted that no further information is provided in the dataset name section regarding how to request access to the mentioned dataset.

Technical reproducibility in DL studies can be achieved if the code and dataset are released by the study group. A major recommendation in the [18] report is to emphasize the open availability of AI code since AI research and code are inextricably linked. There is no doubt that this ideal way of reporting can help a broader scientific community to build upon the initial published work. We screened the papers and searched through the PDF files with keywords: "code," "GitHub," and "share" to see if we could find any information on publishing the study's code online. Despite this effort, among 80 studies that were included, only four studies had their code publicly available (5%). This is in line with the results of a recently published article in the radiology AI journal, which reported that 24 out of 218 (11%) studies shared their code with sufficient documentation to be considered reproducible [19]. Another study reported a 21% code-sharing rate in healthcare AI papers published from 2017–2019 [10]. The lower code-sharing ratio for papers published in the JDI journal indicates a high need for a protocol or guideline to make the studies more reproducible.

A quick note is that although sharing code and data can let us achieve technical reproducibility, the results might be achieved by overfitting, so using another dataset may not give show us results in line with our first ones. Another point is that since sharing the code is not always available,

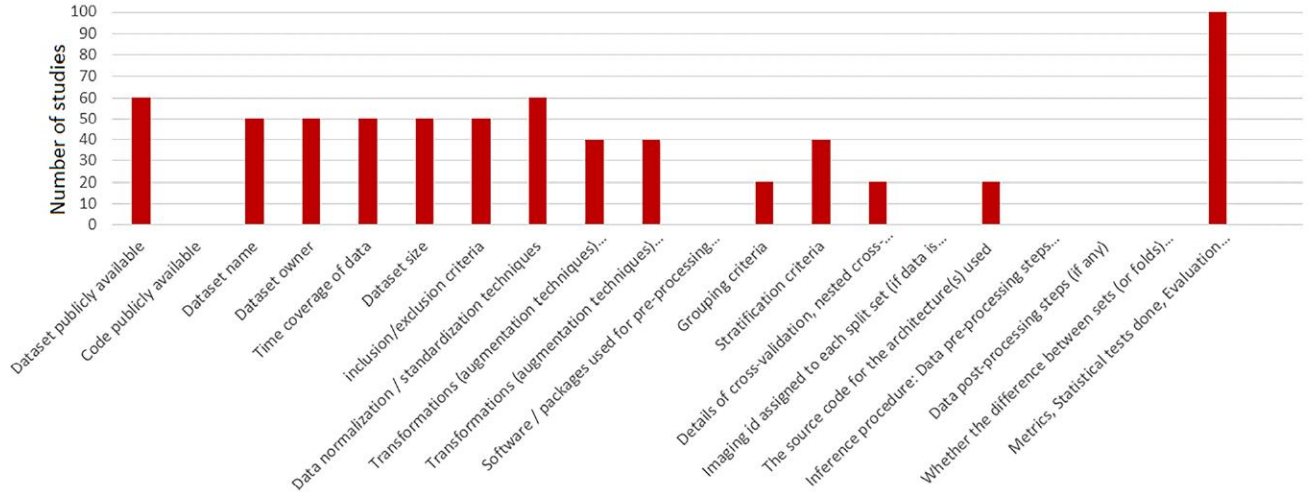

Fig. 5 The illustration of the frequency of the reported items on the external dataset when applicable

sharing a JavaScript Object Notation or Docker file with model details to make it possible for other researchers to reproduce the results is highly helpful.

DL models achieve different results on different datasets. It is, therefore, necessary but not sufficient to have datasets associated with the results for a study to be technically reproducible. However, this matter is not always achievable. It is well-known that data cannot always be shared due to privacy, confidentiality, and security concerns in the context of health-related information. This issue is less problematic in studies that use publicly available datasets. Based on our search, 30 studies (40.5%) used data from open-source datasets, showing the importance of these publicly released datasets to foster reproducible science.

Here, we note that open research, including data and code sharing, does not guarantee reproducibility in all cases. According to a study, it has been found that among the published repositories with papers which happened in 22 of cases, the quality of these repositories is not consistently high [20]. Reproducibility may not be achieved both in terms of statistical replication and technical replication. This can occur due to flawed statistical procedures or, for example, when two researchers utilize different versions of a software library, leading to significant discrepancies in conclusions if essential parameters are assigned different values. Furthermore, even if a paper seems to provide all the necessary details for replication, replication efforts may still fail. One possible reason for such failures could be the involvement of numerous variables in the original study, resulting in a methodology that is specifically applicable to a particular dataset [21, 22].

McDermott et al. reviewed 511 papers presented at machine learning conferences from 2017 to 2019 [10]. Papers were categorized into the following four fields: machine learning applied to health (MLH), natural language processing, computer vision, and general machine learning models. We found no exact criteria reported for this categorization in the paper. They reported that ~55% of the MLH studies used public datasets compared to more than 90% of computer vision and natural language processing papers and ~85% of general machine learning papers. In addition, ~21% of the MLH papers released their code publicly, compared to ~39% of the papers in computer vision and ~48% of the papers in natural language processing. McDermott et al. reported that health-related datasets tend to be relatively small, have high dimensionality, are noisy, and often suffer from sparse/irregular sampling. These common issues related to health-related datasets negatively impact the reproducibility of the associated results.

Wright et al. published a paper investigating reproducibility in radiology papers in general. They measured the reporting of statements regarding open access, funding, conflict of interest, data availability, pre-registration, protocol, analysis scripts, and material availability [23]. This cross-sectional investigation found that the key transparency and reproducibility-related factors were rare or entirely absent among the sample of publications in the field of radiology. None of the analyzed publications reported an analysis script. Furthermore, few of those publications provided access to materials, or few were pre-registered. Only one of those publications provided raw data.

Our review was constrained by the fact that we only examined studies on DL tools that were published in JDI over the past few years. JDI was selected as it is a specialized journal that focuses on publishing papers related to AI and medical imaging. Due to the large number of items reviewed to ensure reproducibility, it was not practical to include publications from all journals. We therefore limited our search to a specific time frame of 2020 to 2022 and conducted a thorough analysis of a limited number of recently published studies.

According to our cross-sectional study, in JDI publications on DL in medical imaging, authors infrequently report the key elements of their study to make it reproducible.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10278-023-00870-5.

Author Contribution All authors contributed to the study conception and design. Material preparation, and analysis were performed mostly by Shahriar Faghani, Mana Moassefi, and data collection was performed by the team. The first draft of the manuscript was written by Mana Moassefi and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Declarations

Ethics Approval This paper is a review. We have not received any approval from universities for conducting this.

Consent to Participate Not applicable.

Consent to Publish Not applicable.

Conflict of Interest The authors declare no competing interests.

References

1. West E, Mutasa S, Zhu Z, Ha R. Global Trend in Artificial Intelligence-Based Publications in Radiology From 2000 to 2018. AJR Am J Roentgenol. 2019;213: 1204–1206.
2. Kelly BS, Judge C, Bollard SM, Clifford SM, Healy GM, Aziz A, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol. 2022. https://doi.org/10.1007/s00330-022-08784-6
3. Oppenheimer J, Lüken S, Hamm B, Niehues SM. A Prospective Approach to Integration of AI Fracture Detection Software in Radiographs into Clinical Workflow. Life. 2023;13. https://doi.org/10.3390/life13010223

4. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, Suever JD, Geise BD, Patel AA, et al. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. NPJ Digit Med. 2018;1: 9.
5. Akkus Z, Cai J, Boonrod A, Zeinoddini A, Weston AD, Philbrick KA, et al. A Survey of Deep-Learning Applications in Ultrasound: Artificial Intelligence-Powered Ultrasound for Improving Clinical Workflow. J Am Coll Radiol. 2019;16: 1318–1328.
6. Yu K-H, Lee T-LM, Yen M-H, Kou SC, Rosen B, Chiang J-H, et al. Reproducible Machine Learning Methods for Lung Cancer Detection Using Computed Tomography Images: Algorithm Development and Validation. J Med Internet Res. 2020;22: e16709.
7. Gundersen OE, Kjensmo S. State of the Art: Reproducibility in Artificial Intelligence. AAAI. 2018;32. https://doi.org/10.1609/aaai.v32i1.11503
8. Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. NPJ Digit Med. 2019;2: 2.
9. McDermott MBA, Wang S, Marinsek N, Ranganath R, Ghassemi M, Foschini L. Reproducibility in Machine Learning for Health. arXiv [cs.LG]. 2019. Available: http://arxiv.org/abs/1907.01463
10. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: Still a ways to go. Sci Transl Med. 2021;13. https://doi.org/10.1126/scitranslmed.abb1655
11. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? Sci Transl Med. 2016;8: 341ps12.
12. Campbell DT. Relabeling internal and external validity for applied social scientists. New Dir Prog Eval. 1986;1986: 67–77.
13. Moher D. Reporting guidelines: doing better for readers. BMC Med. 2018;16: 233.
14. Shelmerdine SC, Arthurs OJ, Denniston A, Sebire NJ. Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. BMJ Health Care Inform. 2021;28. https://doi.org/10.1136/bmjhci-2021-100385
15. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiol Artif Intell. 2020;2: e200029.
16. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc. 2020;27: 2011–2015.
17. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med. 2018;169: 467–473.
18. Kitamura FC, Pan I, Kline TL. Reproducible Artificial Intelligence Research Requires Open Communication of Complete Source Code. Radiol Artif Intell. 2020;2: e200060.
19. Venkatesh K, Santomartino SM, Sulam J, Yi PH. Code and Data Sharing Practices in the Radiology AI Literature: A Meta-Research Study. Radiology: Artificial Intelligence. 2022; e220081.
20. Simko A, Garpebring A, Jonsson J, Nyholm T, Löfstedt T. Reproducibility of the Methods in Medical Imaging with Deep Learning. arXiv [cs.LG]. 2022. Available: http://arxiv.org/abs/2210.11146
21. Beam AL, Manrai AK, Ghassemi M. Challenges to the Reproducibility of Machine Learning Models in Health Care. JAMA. 2020;323: 305–306.
22. Colliot O, Thibeau-Sutre E, Burgos N. Reproducibility in machine learning for medical imaging. arXiv [cs.CV]. 2022. Available: http://arxiv.org/abs/2209.05097
23. Wright BD, Vo N, Nolan J, Johnson AL, Braaten T, Tritz D, et al. An analysis of key indicators of reproducibility in radiology. Insights Imaging. 2020;11: 65.