Black Box or Open Science? A study on reproducibility in AI Development Papers

Florian Königstorfer
BANDAS Center, University of Graz
florian.koenigstorfer@uni-graz.at

Armin Haberl
BANDAS Center, University of Graz
armin.haberl@uni-graz.at

Dominik Kowald
Know-Center & TU Graz
dkowald@know-center.at

Tony Ross-Hellauer
Know-Center & TU Graz
tross@know-center.at

Stefan Thalmann
BANDAS Center, University of Graz
stefan.thalmann@uni-graz.at

Abstract

The surge in Artificial Intelligence (AI) research has spurred significant breakthroughs across various fields. However, AI is known for its Black Box character and reproducing AI outcomes is a challenge. Open Science, emphasizing transparency, reproducibility, and accessibility, is crucial in this context, ensuring research validity and facilitating practical AI adoption. We propose a framework to assess the quality of AI documentation and assess 51 papers. We conclude that despite guidelines, many AI papers fall short on reproducibility due to insufficient documentation. It is crucial to provide comprehensive details on training data, source code, and AI models, and for reviewers and editors to strictly enforce reproducibility guidelines. A dearth of detailed methods or inaccessible source code and models can raise questions about the authenticity of certain AI innovations, potentially impeding their scientific value and their adoption. Although our sample size inhibits broad generalization, this study nonetheless offers key insights on enhancing AI research reproducibility.

Keywords: AI documentation, Reproducibility, Open Science.

1. Introduction

In both science and practice, Artificial Intelligence (AI) appears to lead to significant performance increases compared to more traditional methods (Kapoor & Narayanan, 2022). Despite the well-documented claims of performance improvements, current research indicates that many papers contain significant analytical errors that alter the papers' outcomes once corrected (Kapoor & Narayanan, 2022). Researchers found that some mistakes are so substantial that the developed AIs were not usable for practical purposes (Roberts et al., 2021).

The reasons why expert readers frequently struggle to identify errors in scientific work include not only the unpredictable behavior of some AI models, but also the fact that vital information pertaining to the experiment, such as data details, and the code utilized for the experiment, are frequently omitted (Gundersen & Kjensmo, 2018). Research in the area of "Open Science" (OS) has shown that making research transparent can help reduce mistakes made by researchers (Smaldino et al., 2019), and help ensure reproducibility. To minimize errors and enhance reproducibility in science, checklists and guidelines for documenting information concerning the development of AI have been shown to be useful (Collins et al., 2015; Kapoor & Narayanan, 2022; Mitchell et al., 2019).

Lack of reproducibility is a particularly significant problem for papers focusing on developing AI. Applying poorly built or poorly understood AI models without being able to investigate potential errors can cause significant harm in practice (Estiri et al., 2022; Saisubramanian et al., 2022). However, despite studies having reviewed the documentation quality of papers that use AI to analyze and interpret data patterns (Gundersen & Kjensmo, 2018) there is still no review of the quality of documentation for papers on AI development. Such a review is important since the documentation can help to make papers more reproducible, identify errors in AI development, and ensure that AI can be adopted in practice. The research question is:

Are scientific papers focusing on the development of AI documented well enough to reproduce them?

To answer these questions, we systematically evaluate factors linked to reproducibility in research papers from three domains: information systems (IS), computer science (CS), and medicine. We find that many AI reviewed papers lack sufficient documentation, making their research non-reproducible.

2. Background

AI and the reproducibility problem. Recently, AI has gained attention from researchers and practitioners, showing considerable positive impacts on commercial banks, audit firms, and other industries (Königstorfer & Thalmann, 2020; Krieger et al., 2021). However, many published and peer-reviewed AIs are significantly flawed, compromising their outcomes and practical value (Kapoor & Narayanan, 2022; Roberts et al., 2021). Detecting errors in science is challenging due to two main factors. First, AI derives its decision-making from training data, not developer instructions (Gebru et al., 2021), making it sensitive to biases, noise, and errors in the data. This Black Box nature of AI (Castelvecchi, 2016) obscures the reasoning behind decisions. Second, research shows that flawed results often stem from experimental errors and biases in AI models (Kapoor & Narayanan, 2022; Roberts et al., 2021), including the absence of test sets, use of invalid features, and datasets with duplicates (Arp et al., 2020; Roberts et al., 2021).

However, even glaring errors and biases are often not caught. One reason behind the inability to detect these errors is that the implementation of scientific studies is hard to reproduce since essential aspects of the implementation, including data, experimental setup, and research question, are not described and inaccessible (Gundersen & Kjensmo, 2018). As a result, errors in research persist. Thus, it remains unclear whether papers on AI development can be adequately reproduced. Given that AI models may be applied in practical settings in their current or slightly modified form, academic literature concerning their development must adhere to a more rigorous standard.

Open Science, and documentation of AI. Such rigorous standards for science in general are defined in the OS community, involving transparent, accessible knowledge shared and developed via collaborative networks (Vicente-Saez & Martinez-Fuentes, 2018). OS focuses on components like Open Access, Open Data, Open Methodology, Open Artifacts, and Open Peer Review. For reproducibility, which is replicating results independently using the same AI method based on original documentation (Gundersen & Kjensmo, 2018), Data, Open Methodology, and Open Artifacts are essential. Open Data disseminates data freely, Open Methodology involves detailed method documentation, and Open Artifacts are accessible research artifacts with an open license (Abele-Brehm et al., 2019; Doyle et al., 2019; Jomier, 2017).

Achieving reproducibility in AI research requires thorough documentation. In software engineering, "documentation" refers to maintaining a clear and accessible record of architectural decisions (Clements et al., 2011). This record captures essential software components. For AI applications, documentation must meet specific requirements to ensure reproducibility (Gundersen & Kjensmo, 2018).

Various methods and tools exist for AI documentation, including data, model, and field-specific documentation methods (Gebru et al., 2021; Mitchell et al., 2019; Rivera et al., 2020). These tools are essential for transparency and research integrity, helping to meet Open Data, Open Methodology, and Open Artifacts standards. Table A.1 in the digital appendix lists tools for creating open, reproducible AI models. Tools and documentation are critical, as ill-constructed AI models can be harmful, particularly if errors are not scrutable (Estiri et al., 2022; Saisubramanian et al., 2022).

OS offers multiple advantages, including allowing the use of existing data and code and enabling research evaluation even without full replicability (Peng, 2011). Transparent practices also minimize errors since data, code, and methodology transparency enables faster error detection (Gundersen & Kjensmo, 2018; Smaldino et al., 2019). This yields higher quality research and effective AI applications in practice. However, due to the characteristics of AI and in particular due to the Black Box character, it is unclear whether academic studies are documented well enough to be reproduced.

3. Methodology

Our analysis aims to investigate (1) how well researchers document the AI development in their studies and (2) in which areas of OS the documentation needs to be improved. To this end, a literature review was conducted in the top journals in the fields of *IS*, *CS* and Medicine. To evaluate how well the papers were documented, we investigate whether the papers contain the information necessary to reproduce the results. This information is based on an extended version of the reproducibility framework presented by Heil et al. (2021). We expect that using this methodology, we will be able to tell whether the reviewed papers focused on AI development in top journals in the fields of *IS*, *CS* and Medicine are sufficiently well documented to be reproduced by experts. Even though we are not striving for broad generalizability, this exploratory study gives researchers a better understanding of how to improve the reproducibility of their studies.

Data collection. Identification of relevant literature occurred in two rounds. Initially, papers were collected from Senior Scholars' List of Premier Journals and three renowned international conferences (ECIS, ICIS, HICSS) through a Google Scholar keyword search, considering papers published post-2010. Despite challenges with Google Scholar (Pieper et al., 2021), it remains a reputable engine for AI-related scientific publications (Gray et al., 2012). Keywords were defined via Rowley and Slack's (2004) building block approach. The terms included "data", "algorithm", "code", "accuracy", and "artificial intelligence", or "machine learning". The search excluded papers including the term "survey" in the title.

From 1,042 search results, we included only relevant papers focused on AI development with the goal of solving a problem. Within the field of medicine, for example, this could be an AI model that predicts whether a patient has a specific disease. Excluded were papers that did not develop an AI (i.e. papers about ethical/responsible AI discussions, surveys, new benchmark datasets, and new methods for preparing data) and papers that did not explicitly solve a problem (i.e. papers presenting optimized neural architectures for specific datasets). From the title, it was, in many cases, clear that the papers did not develop any AI. Titles indicating AI development were considered for inclusion in this study and evaluated in the next step, while others pointing at exclusions were rejected. Next, abstract scans were conducted. Papers highlighting AI development were included, while empirical studies (studies in which the data analysis using Machine Learning (ML) is the main contribution), non-English papers or inaccessible ones were discarded. This process, conducted Q1 2023, resulted in 17 reviewed studies.

To expand the number of studies in this review, a second round sourced from the top 18 journals and conferences in the "Artificial Intelligence" sub-ranking of the Journal Citation Report ranked by 2021 JIF was conducted. Again, keywords were defined using the building block approach of Rowley & Slack (2004). The keywords "data", "algorithm", "code", "accuracy", and "artificial intelligence" were used. In addition to that, any papers containing the terms "survey" or "framework" in the title were excluded. This search was also restricted to post-2010 publications, resulting in 4257 search results. The title and abstract scan followed the same inclusion and exclusion criteria as in the first round. After the title review and abstract scan, this round (Q1 2023), 34 additional papers were included. In total, 51 papers were analyzed in this study.

To obtain a more detailed view of the results, the papers were categorized into one of three groups: IS papers, CS papers, and Medical papers. The categorization was primarily based on the classification of the journal presented in the Journal Citation Report 2021. However, the search for papers in one journal that was classified as a Computer Science journal (Nature Machine Intelligence) resulted primarily in Medical papers. Due to this issue, the classification of the papers from this journal was based on the content of the individual papers. Table 1 presents the number of papers in each category.

Category	Number of papers
Computer Science	22
Information Systems	17
Medicine	12

Table 1: number of papers in each category.

A detailed list of journals and conferences, as well as the number of papers from each journal and conference, can be found in Table A.2 in the digital appendix. Table A.4 in the digital appendix presents the search terms for Google Scholar.

Analysis. This paper analyzes the documentation quality of AI in literature using a framework extended from Heil et al.'s (2021) reproducibility standard. Heil et al. (2021) present an evaluation framework for categorizing papers focused on AI development into reproducibility categories based on documentation details. Based on previous research, we integrated three additional questions concerning data collection, cleaning, and feature engineering; suitable application cases; and AI model performance (Königstorfer & Thalmann, 2021). These questions create a checklist to evaluate transparency and reproducibility by assessing the availability of data, models, and source code availability (Q1-3), installation commands (Q4), analysis details (Q5), determinism settings (Q6), and reproducibility in a single command (Q7). They also assess practical relevance through application cases (Q8) and information depth on data processing (Q9) and AI performance (Q10).

For Bronze classification, a paper must address questions Q1-3. For Silver, it must also include Q4-6. Gold classification requires satisfactory answers to all questions. Table A.3 in the digital appendix outlines the questions and categorization. The reproducibility framework by Heil et al. (2021) was chosen for its ease in aligning questions with different OS components. For example, Q1 relates to "Open Data," while Q4 to Q7 and Q9 address "Open Methodology".

Papers were assessed by one author against questions in Table A.3, confirming the presence of required information. If available, the question was answered "Yes". Even with non-working source code links like in Wang et al. (2020), based on journal retention standards (Joly et al., 2012), a "Yes" was given. Referenced but limited information got a "Partial" response. This answer also included papers with non-shareable medical data (Niu et al., 2019). A "No" was given if no references or access were provided. A second author, blind to the first's findings, reviewed five papers and tried to answer the Table A.3 questions. Agreement was initially 82%, but after discussion, it increased to 96%. The team resolved the remaining discrepancies through discussion. The analysis was done by knowledgeable experts in the field of AI and ML.

4. Results

Overall reproducibility of the papers. Table 2 presents the percentage of papers from each academic field in each reproducibility category defined in Table A.3.

Label	Percentage of CS papers	Percentage of IS papers	Percentage of Medicine papers
no label	77,3 %	94,1 %	75 %
Bronze	13,7 %	5,9 %	25 %
Silver	9 %	0 %	0 %
Gold	0 %	0 %	0 %

Table 2: overview of results.

Table 2 reveals that no label was assigned to most reviewed papers. These papers often exclude references to training data, source code, and AI models. Only two papers provide sufficient information for the "Silver" label (Agostinelli et al., 2019; Xu & Carpuat, 2021), and none reach the "Gold" standard. One paper nearly achieves Gold (Agostinelli et al., 2019), but lacks information on a potential integration of the AI into an organizational setting.

Many papers without label omit basic elements like data, source code, and AI model access. Approximately a third of all papers offer neither access nor references to any of these components without giving reasons (Abbasi et al., 2019; Ben-Assuli & Padman, 2020; Chau et al., 2020; Chen et al., 2022; Choi et al., 2019; Khan et al., 2020; López-Linares et al., 2018; Munnangi & Paruchuri, 2020; Shi et al., 2020; Singh & Tucker, 2017; Teuwen et al., 2021; Tofangchi et al., 2017; Wu et al., 2021; Xue et al., 2021; Yet et al., 2013; B. Zhang et al., 2018). Some provide two of the requested items (Ferreira et al., 2016; Niu et al., 2019; Pu et al., 2018; Twitchell & Fuller, 2019; Valvoda et al., 2022; Wang et al., 2020; H. Yang et al., 2022), while all others meet only a single "Bronze" requirement. Among "Bronze" papers, all but two lacked information about deterministic analysis and single-command dependency installation.

Across the three fields, the reviewed papers show low reproducibility, with most not reaching "Bronze". Differences are seen in "Bronze" paper distribution: 25% of medical and 22.7% of *CS* papers achieve "Bronze" or higher, versus 5.88% of *IS* papers. Only two *CS* papers offer sufficient access and information for "Silver" (Agostinelli et al., 2019; Xu & Carpuat, 2021), indicating limited support for reproducibility and practical AI usage information for most papers in this review.

Reproducibility of the OS aspects. A closer examination of OS's components provides a nuanced response to the research question. Tables 3 to 5 show significant disparities in documentation across various academic fields.

Traceability of the data. Our analysis shows that 81,9% of *CS*, 47 % of *IS*, 66,7% and of Medicine papers make the data at least partially available. However, data accessibility is limited by several factors. An issue particularly prevalent in medical research is the citation of data, in combination with access restrictions due to privacy concerns (Buettner et al., 2019; Das et al., 2021; Deshpande et al., 2015; Niu et al., 2019; C. Yang et al., 2015; Yun et al., 2019). Occasionally, authors' download links are invalid (Das et al., 2021; Wang et al., 2020) or lead to empty online storage folders (Yu et al., 2021). *IS* papers often use publicly available data (Twitter data or legal texts) but do not provide necessary labels for analysis (Abbasi et al., 2019; Michel et al., 2022).

The inability to access certain datasets is a significant issue, even with a well-documented data collection process. Papers utilizing social media data often indicate sources and collection times but fail to share the dataset (Abbasi et al., 2019). This is problematic, since it is nearly impossible to replicate the exact dataset, even with identical search parameters. Social media users might delete their posts or account, face bans, or the platform could cease operations, resulting in data loss.

Reproducibility of the methodology. Table 3 shows the methodology information provided by the reviewed papers in each field.

Question	Percentage of *CS* papers	Percentage of *IS* papers	Percentage of Medicine papers
(Q 4)	18,1 %	5,9 %	8,3 %

(Q 5)	100 %	100 %	100 %
(Q 6)	13,6 %	0 %	8,3 %
(Q 7)	4,5 %	0 %	8,3 %
(Q 9)	100 %	100 %	100 %

Table 3: description of the methodology.

Table 3 reveals that reproducing the papers' using the information on methodology provided inside the papers is tough. Few facilitate easy dependency installation. Providing a Docker file containing source code, dependencies, and AI model is considered best practice but is rarely implemented (Agostinelli et al., 2019; Di He et al., 2022). However, most papers share key analysis details such as computational resources, operating system, AI algorithm, hyperparameters, and programming language (Buettner et al., 2019; Choi et al., 2019; Das et al., 2021; Jeon et al., 2021; Liu et al., 2020; López-Linares et al., 2018; Pu et al., 2018; Valvoda et al., 2022; Yu et al., 2021). A few document only one component or less (Tofangchi et al., 2017; Zhou et al., 2022). Many omit the operating system used for AI model training. Only four specify deterministic analysis, crucial for "Gold" or "Silver" categorization. Besides one, none allow full reproducibility with a single command (Alaa et al., 2021). Some papers, though not reproducible in one command, offer training data and source code for various stages of the development process.

Table 3 also indicates that while all papers provide at least some data collection, cleaning, and feature engineering information, the detail varies. Few cover all three aspects thoroughly (Chau et al., 2020; Chen et al., 2022; Liu et al., 2020; Munnangi & Paruchuri, 2020; Pu et al., 2018; Tofangchi et al., 2017; Twitchell & Fuller, 2019; Zhou et al., 2022), others share limited or no information (Badrinath et al., 2016; Choi et al., 2019; Ho et al., 2020; Quellec et al., 2021; Wu et al., 2021; C. Yang et al., 2015; B. Zhang et al., 2018; L. Zhang et al., 2020). However, missing data collection details are not always critical, when data originates from government agencies or is collected by certified machines and verified by professionals, assuming higher data quality as a result (Agostinelli et al., 2019; Ben-Assuli & Padman, 2020; Jeon et al., 2021; Munnangi & Paruchuri, 2020; Niu et al., 2019; Quellec et al., 2021; Shi et al., 2020; C. Yang et al., 2015).

Regarding information on the feature engineering, its absence may not be a major concern when using neural networks, as they inherently extract features from training data (Antelis et al., 2020; Ho et al., 2020; Khan et al., 2020; Michel et al., 2022; Quellec et al., 2021; Teuwen et al., 2021; Wang et al., 2020; Yu et al., 2021; Yun et al., 2019). However, the lack of data cleaning details is concerning and can impact model performance notably (Abbasi et al., 2019; Teuwen et al., 2021; Yu et al., 2021; Yun et al., 2019). This issue is amplified in studies using AI algorithms sensitive to outliers and other errors without data cleaning information. Several of the reviewed papers that used convolutional neural networks and artificial neural networks fell short in disclosing their data cleaning process (Teuwen et al., 2021; Yu et al., 2021; Yun et al., 2019).

Accessibility of the artifacts. Table 4 displays the percentage of papers in each field that provide at least partial openness to their artifacts.

Question	Percentage of *CS* papers	Percentage of *IS* papers	Percentage of Medicine papers
(Q 2)	22,7%	5,9%	41,7 %
(Q 3)	40,9%	11,8%	41,7 %
(Q 4)	18,1 %	5,9%	8,3%
(Q 8)	9 %	70,5%	25 %
(Q 10)	100 %	94,1%	100%

Table 4: information on the artifacts.

Table 4 reveals that the majority of reviewed papers across all fields lack openness concerning their artifacts. Most papers neither grant public access nor reference the source code for training the AI or the final AI models (Q 2 & 3). Roughly one-third of all papers reviewed do not provide access or references to their data, source code, or final AI model (Abbasi et al., 2019; Ben-Assuli & Padman, 2020; Chau et al., 2020; Chen et al., 2022; Choi et al., 2019; Khan et al., 2020; López-Linares et al., 2018; Munnangi & Paruchuri, 2020; Shi et al., 2020; Singh & Tucker, 2017; Teuwen et al., 2021; Tofangchi et al., 2017; Wu et al., 2021; Xue et al., 2021; Yet et al., 2013; B. Zhang et al., 2018). However, a few papers offer access to a Docker file including the source code or model and related software dependencies (Q 2, 3 & 4) (Agostinelli et al., 2019; Di He et al., 2022). Notably, all but one paper detail the AI application's performance (Q 10), providing metrics such as accuracy, AUC, cross-validation performance, etc.

Intriguingly, some papers lacking a direct source code or model link still had repositories on GitHub citing them (Niu et al., 2019; Yu et al., 2021). In one case, a GitHub search of the paper title revealed three repositories with identical titles (Yu et al., 2021). These repositories were created following the paper's instructions, but their origin (paper's authors or a third party) often remains unclear.

Most *CS* and medicine papers neglect practical applications or governance details of their AI models, raising concerns, especially given AI's potential impact in medicine. However, some *IS* research provides examples of addressing potential applications and misuse risks (Abbasi et al., 2019; Valvoda et al., 2022).

5. Discussion

This paper has shown that of the literature assessed, most papers provide little documentation, leading to weak reproducibility. These results are broadly in line with the earlier findings of (Gundersen & Kjensmo, 2018). Hence, despite our relatively small sample of papers assessed, we believe this evidence further demonstrates significant issues of transparency of documentation across the AI literature. Despite AI research guidelines (Gebru et al., 2021; Mitchell et al., 2019; Rivera et al., 2020), and specific instructions from journals and conferences (AAAI, 2023), many papers do not follow the guidelines closely. Researchers need to furnish more information on training data, source code, and AI models, while reviewers and editors should strictly enforce these guidelines.

Analyzing OS's components reveals different patterns. Most papers cite or provide access to training data, but some limit access to their data. Research on public benchmark datasets such as ImageNet has shown many errors, invalidating numerous studies using them (Northcutt et al., 2021b). This shows that access and references to the used data is particularly important. Since tools to identify and correct the errors mentioned in the previous paragraph are available (Northcutt et al., 2021a), it is also essential to know whether these tools were used and how the data was cleaned and prepared. However, many papers lack comprehensive methodological information. The lack of information on the data cleaning process, combined with existing research revealing the impact of errors in the data cleaning process on the results (Kapoor & Narayanan, 2022; Northcutt et al., 2021b) raises doubts about the results reported in several papers. Also, given the reliance of AI on clean and unbiased training data (Gebru et al., 2021), the lack of information on these aspects makes it hard for technical experts and reviewers to reproduce the papers. Furthermore, many papers do not present information on potential AI's, governance and security processes, and AI integration into business and medical processes. This information is essential for practical AI applications. For example, one study reportedly spent about 1.3 million euros on Cloud Computing for training AI models (L. Zhang et al., 2020). Information on (1) whether this financial investment can be recouped and (2) how users need to be protected from potential damages is essential. This is a particular challenge for the adoption of the proposed AIs in practice, due to recent EU and US regulations that require documentation and details on AI's integration into business and associated risk management systems (European Commission, 2021; NIST, 2022).

While all three fields have similar reproducibility for methodology, they differ in data and artifact accessibility. IS literature offers limited data, source code, and AI model access but details AI integration into organizations. Conversely, CS and medicine provide minimal guidance on AI application, governance, and integration but excel in describing the data and model. The superior documentation in CS and medicine compared to IS might be due to existing guidelines and field norms. Documentation standards for AI in medicine and CS have been proposed (Bender & Friedman, 2018; Rivera et al., 2020), while IS standards lag behind, leading to inadequate AI documentation and challenges to the adoption in practice (Königstorfer & Thalmann, 2020). However, we have not deeply explored if guidelines directly improve documentation quality in the disciplines. Further research is needed.

A potential reason for non-reproducibility might be the scant guidelines from academic outlets. Many esteemed academic platforms do not mandate authors to share datasets, AI frameworks, or codes during publication. For example, in the Senior Scholars' List of Premier Journals only the Information Systems Journal (2023) requires disclosure of training data and AI models. This raises concerns about the scientific validity of publications in the other journals of the Senior Scholars' List of Premier Journals. Future studies should evaluate journal standards and their adherence in AI articles. Reviewing the validity of non-compliant publications is also essential.

The importance of reproducibility depends on the context in practice as well as in science. In practice, the EU's draft of the AI Act defines this context based on a risk approach (European Commission, 2021). For high risk use cases a sound documentation and reproducibility will soon become necessary. In science, we currently start this differentiation and we see that in high risk fields, like medicine, we have more advanced standards. However, this is a promising topic for the OS community.

In conclusion, AI documentation in the reviewed scientific publications is subpar and requires enhancement to meet OS standards. Our study initiates awareness and introduces an evaluation scheme for assessing documentation quality. Future research should develop domain-specific documentation standards, augmented by technical reproducibility studies of published papers. A compelling research direction is whether people with different levels of expertise can replicate the paper's results with the given information. For example, Automated Machine Learning (AutoML) might assist in developing and comparing AI models and data preparation techniques (Polzer & Thalmann, 2022).

6. Conclusion

In conclusion, this paper has highlighted the inadequate level of reproducibility in the majority of the reviewed AI research papers. Even though our sample size limits broad generalization, our study provides vital insights into improving AI research reproducibility, especially considering many researchers often omit crucial details like training data, source code, and AI models. To ensure the credibility and validity of AI research, it is imperative that researchers provide more detailed information on their methodology and data, while reviewers and editors should enforce guidelines more strictly. The provision of research data from papers is critical for identifying common errors, and promoting reuse by other researchers in their studies. Overall, science needs more research on reproducibility guidelines and standards for AI research.

Acknowledgements: this research was co-funded by the European Union (Horizon Europe project TIER2, GA: 101094817).

7. References

AAAI. (2023). *Reproducibility Checklist.* Association for the Advancement of Artificial Intelligence (AAAI). https://aaai.org/conference/aaai/aaai-23/reproducibility-checklist/

Abbasi, A., Li, J [Jingjing], Adjeroh, D., Abate, M., & Zheng, W. (2019). Don't mention it? Analyzing user-generated content signals for early adverse event warnings. *Information Systems Research*, *30*(3), 1007–1028.

Abele-Brehm, A. E., Gollwitzer, M., Steinberg, U., & Schönbrodt, F. D. (2019). Attitudes toward open science and public data sharing. *Social Psychology*.

Agostinelli, F., McAleer, S., Shmakov, A., & Baldi, P. (2019). Solving the Rubik's cube with deep reinforcement learning and search. *Nature Machine Intelligence*, *1*(8), 356–363.

Alaa, A. M., Gurdasani, D., Harris, A. L., Rashbass, J., & van der Schaar, M. (2021). Machine learning to guide the use of adjuvant therapies for breast cancer. *Nature Machine Intelligence*, *3*(8), 716–726.

Antelis, J. M., Falcón, L. E., & others (2020). Spiking neural networks applied to the classification of motor tasks in EEG signals. *Neural Networks*, *122*, 130–143.

Arp, D., Quiring, E., Pendlebury, F., Warnecke, A., Pierazzi, F., Wressnegger, C., Cavallaro, L., & Rieck, K. (2020). Dos and don'ts of machine learning in computer security. *ArXiv Preprint ArXiv:2010.09470*.

Badrinath, N., Gopinath, G., Ravichandran, K. S., & Soundhar, R. G. (2016). Estimation of automatic detection of erythemato-squamous diseases through adaboost and its hybrid classifiers. *Artificial Intelligence Review*, *45*(4), 471–488.

Ben-Assuli, O., & Padman, R. (2020). Trajectories of Repeated Readmissions of Chronic Disease Patients: Risk Stratification, Profiling, and Prediction. *Mis Quarterly*, *44*(1).

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604.

Buettner, R., Frick, J., & Rieg, T. (2019). High-performance detection of epilepsy in seizure-free EEG recordings: A novel machine learning approach using very specific epileptic EEG sub-bands. In *ICIS*.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, *538*(7623), 20.

Chau, M., Li, T. M. H., Wong, P. W. C., Xu, J. J., Yip, P. S. F., & Chen, H. (2020). Finding People with Emotional Distress in Online Social Media: A Design Combining Machine Learning and Rule-Based Classification. *Mis Quarterly*, *44*(2).

Chen, Z [Zhi], Liu, Y [Yuncong], Chen, L., Zhu, S., Wu, M [Mengyue], & Yu, K. (2022). OPAL: Ontology-Aware Pretrained Language Model for End-to-End Task-Oriented Dialogue. *ArXiv Preprint ArXiv:2209.04595*.

Choi, J., Youn, J.-H., & Haas, C. (2019). Machine Learning Approach for Foot-side Classification using a Single Wearable Sensor. In *ICIS*.

Clements, P., Bachmann, F., Bass, L., Garlan, D., Ivers, J., Little, R., Merson, P., Nord, R., & Stafford, J. (2011). *Documenting Software Architectures: Views and Beyond*. Addison-Wesley Professional.

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Annals of Internal Medicine*, *162*(1), 55–63.

Das, P. K., Meher, S., Panda, R., & Abraham, A. (2021). An efficient blood-cell segmentation for the detection of hematological disorders. *IEEE Transactions on Cybernetics*, *52*(10), 10615–10626.

Deshpande, G., Wang, P., Rangaprakash, D., & Wilamowski, B. (2015). Fully connected cascade artificial neural network architecture for attention deficit hyperactivity disorder classification from functional magnetic resonance imaging data. *IEEE Transactions on Cybernetics*, *45*(12), 2668–2679.

Di He, Liu, Q., Wu, Y., & Xie, L. (2022). A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening. *Nature Machine Intelligence*, *4*(10), 879–892.

Doyle, C., Luczak-Roesch, M., & Mittal, A. (2019). We need the open artefact: Design Science as a pathway to Open Science in Information Systems research. In *Extending the Boundaries of Design Science Theory and Practice: DESRIST 2019, Worcester, MA, USA, June 4‑6, 2019, Proceedings 14.* Symposium conducted at the meeting of Springer.

Estiri, H., Strasser, Z. H., Rashidian, S., Klann, J. G., Wagholikar, K. B., McCoy Jr, T. H., & Murphy, S. N. (2022). An objective framework for evaluating unrecognized bias in medical AI models predicting COVID-19 outcomes. *Journal of the American Medical Informatics Association*, *29*(8), 1334–1341.

European Commission. (2021). *Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts*. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

Ferreira, A., Felipussi, S. C., Alfaro, C., Fonseca, P., Vargas-Munoz, J. E., Dos Santos, J. A., & Rocha, A. (2016). Behavior knowledge space-based fusion for copy‑move forgery detection. *IEEE Transactions on Image Processing*, *25*(10), 4729–4742.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12).

Gray, J. E., Hamilton, M. C., Hauser, A., Janz, M. M., Peters, J. P., & Taggart, F. (2012). Scholarish: Google Scholar and its value to the sciences. *Issues in Science and Technology Librarianship*, *70*(Summer).

Gundersen, O. E., & Kjensmo, S. (2018). State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Heil, B. J., Hoffman, M. M., Markowetz, F., Lee, S.‑I., Greene, C. S., & Hicks, S. C. (2021). Reproducibility standards for machine learning in the life sciences. *Nature Methods*, 1–4.

Ho, C.‑J., Chan, C.‑C., & Chen, H. H. (2020). AF-Net: A convolutional neural network approach to phase detection autofocus. *IEEE Transactions on Image Processing*, *29*, 6386–6395.

Information Systems Journal. (2023). *Guide for authors*. https://onlinelibrary.wiley.com/page/journal/13652575/homepage/forauthors.html

Jeon, S., Choi, W., Park, B., & Kim, C. (2021). A deep learning-based model that reduces speed of sound aberrations for improved in vivo photoacoustic imaging. *IEEE Transactions on Image Processing*, *30*, 8773–8784.

Joly, Y., Dove, E. S., Kennedy, K. L., Bobrow, M., Ouellette, B. F., Dyke, S. O. M., Kato, K., & Knoppers, B. M. (2012). Open science and community norms: data retention and publication moratoria policies in genomics projects. *Medical Law International*, *12*(2), 92–120.

Jomier, J. (2017). Open science‑towards reproducible research. *Information Services & Use*, *37*(3), 361–367.

Kapoor, S., & Narayanan, A. (2022). Leakage and the reproducibility crisis in ML-based science. *ArXiv Preprint ArXiv:2207.07048*.

Khan, M. A., Kwon, S., Choo, J., Hong, S. M., Kang, S. H., Park, I.‑H., Kim, S. K., & Hong, S. J. (2020). Automatic detection of tympanic membrane and middle ear infection from oto-endoscopic images via convolutional neural networks. *Neural Networks*, *126*, 384–394.

Königstorfer, F., & Thalmann, S. (2020). Applications of Artificial Intelligence in commercial banks–A research agenda for behavioral finance. *Journal of Behavioral and Experimental Finance*, *27*, 100352.

Krieger, F., Drews, P., & Velte, P. (2021). Explaining the (non-) adoption of advanced data analytics in auditing: A process theory. *International Journal of Accounting Information Systems*, *41*, 100511.

Liu, Y [Yinhan], Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, *8*, 726–742.

López-Linares, K., Aranjuelo, N., Kabongo, L., Maclair, G., Lete, N., Ceresa, M., Garc\'\ia, I., & Ballester, M. A. G. (2018). Fully automatic detection and segmentation of abdominal aortic thrombus in post-operative CTA images using deep convolutional neural networks. *Medical Image Analysis*, *46*, 202–214.

Michel, M., Djurica, D., & Mendling, J. (2022). Identification of Decision Rules from Legislative Documents Using Machine Learning and Natural Language Processing. In *HICSS*.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*.

Munnangi, S. K., & Paruchuri, P. (2020). Improving Wildlife Monitoring using a Multi-criteria Cooperative Target Observation Approach. In *HICSS*.

NIST. (2022). *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf

Niu, X., Shan, S., Han, H [Hu], & Chen, X. (2019). Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, *29*, 2409–2423.

Northcutt, C. G., Athalye, A., & Mueller, J. (2021a). *Label Errors in Benchmark ML Test Sets*. https://github.com/cleanlab/label-errors

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

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227.

Pieper, D., Heß, S., & Faggion, C. M. (2021). A new method for testing reproducibility in systematic reviews was developed, but needs more testing. *BMC Medical Research Methodology*, *21*, 1–8.

Polzer, A. K., & Thalmann, S. (2022). The impact of AutoML on the AI development process. *Proceedings of the 2022 Pre-ICIS SIGDSA Symposium*, *13*.

Pu, X., Pappas, N., Henderson, J., & Popescu-Belis, A. (2018). Integrating weakly supervised word sense disambiguation into neural machine translation. *Transactions of the Association for Computational Linguistics*, *6*, 635–649.

Quellec, G., Al Hajj, H., Lamard, M., Conze, P.-H., Massin, P., & Cochener, B. (2021). ExplAIn: Explanatory artificial intelligence for diabetic retinopathy diagnosis. *Medical Image Analysis*, *72*, 102118.

Rivera, S. C., Liu, X., Chan, A.-W., Denniston, A. K., & Calvert, M. J. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Bmj*, *370*.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., & others (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*(3), 199–217.

Rowley, J., & Slack, F. (2004). Conducting a literature review. *Management Research News*.

Saisubramanian, S., Zilberstein, S., & Kamar, E. (2022). Avoiding negative side effects due to incomplete knowledge of AI systems. *AI Magazine*, *42*(4), 62–71.

Shi, D., Guan, J., Zurada, J. M., & Levitan, A. S. (2020). Improving Prediction Models for Mass Assessment: A Data Stream Approach. In *HICSS*.

Singh, A., & Tucker, C. S. (2017). A machine learning approach to product review disambiguation based on function, form and behavior classification. *Decision Support Systems*, *97*, 81–91.

Smaldino, P. E., Turner, M. A., & Contreras Kallens, P. A. (2019). Open science and modified funding lotteries can impede the natural selection of bad science. *Royal Society Open Science*, *6*(7), 190194.

Teuwen, J., Moriakov, N., Fedon, C., Caballo, M., Reiser, I., Bakic, P., Garc\'\ia, E., Diaz, O., Michielsen, K., & Sechopoulos, I. (2021). Deep learning reconstruction of digital breast tomosynthesis images for accurate breast density and patient-specific radiation dose estimation. *Medical Image Analysis*, *71*, 102061.

Tofangchi, S., Hanelt, A., & Böhrnsen, F. (2017). Distributed Cognitive Expert Systems in Cancer Data Analytics: A Decision Support System for Oral and Maxillofacial Surgery. In *ICIS*.

Twitchell, D. P., & Fuller, C. M. (2019). Advancing the assessment of automated deception detection systems: Incorporating base rate and cost into system evaluation. *Information Systems Journal*, *29*(3), 738–761.

Valvoda, J., Cotterell, R., & Teufel, S. (2022). On the Role of Negative Precedent in Legal Outcome Prediction. *ArXiv Preprint ArXiv:2208.08225*.

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, *88*, 428–436. https://doi.org/10.1016/j.jbusres.2017.12.043

Wang, L., You, Z.-H., Li, J.-Q., & Huang, Y.-A. (2020). IMS-CDA: prediction of CircRNA-disease associations from the integration of multisource similarity information with deep stacked autoencoder model. *IEEE Transactions on Cybernetics*, *51*(11), 5522–5531.

Wu, X., Han, H [Honggui], & Qiao, J. (2021). Data-driven intelligent warning method for membrane fouling. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(8), 3318–3329.

Xu, W., & Carpuat, M. (2021). EDITOR: An edit-based transformer with repositioning for neural machine translation with soft lexical constraints. *Transactions of the Association for Computational Linguistics*, *9*, 311–328.

Xue, W., Cao, C., Liu, J., Duan, Y., Cao, H., Wang, J [Jian], Tao, X., Chen, Z [Zejian], Wu, M [Meng], Zhang, J [Jinxiang], & others (2021). Modality alignment contrastive learning for severity assessment of COVID-19 from lung ultrasound and clinical information. *Medical Image Analysis*, *69*, 101975.

Yang, C., Deng, Z., Choi, K.-S., & Wang, S. (2015). Takagi-Sugeno-Kang transfer learning fuzzy logic system for the adaptive recognition of epileptic electroencephalogram signals. *IEEE Transactions on Fuzzy Systems*, *24*(5), 1079–1094.

Yang, H., Li, J [Jiali], Lim, K. Z., Pan, C., van Truong, T., Wang, Q., Li, K., Li, S., Xiao, X., Ding, M., & others (2022). Automatic strain sensor design via active learning and data augmentation for soft machines. *Nature Machine Intelligence*, *4*(1), 84–94.

Yet, B., Bastani, K., Raharjo, H., Lifvergren, S., Marsh, W., & Bergman, B. (2013). Decision support system for Warfarin therapy management using Bayesian networks. *Decision Support Systems*, *55*(2), 488–498.

Yu, C., Gao, C., Wang, J [Jingbo], Yu, G., Shen, C., & Sang, N. (2021). Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, *129*, 3051–3068.

Yun, J., Park, J., Yu, D., Yi, J., Lee, M., Park, H. J., Lee, J.-G., Seo, J. B., & Kim, N. (2019). Improvement of fully automated airway segmentation on volumetric computed tomographic images using a 2.5 dimensional convolutional neural net. *Medical Image Analysis*, *51*, 13–20.

Zhang, B., Xiong, D., Su, J., & Qin, Y. (2018). Alignment-supervised bidimensional attention-based recursive autoencoders for bilingual phrase representation. *IEEE Transactions on Cybernetics*, *50*(2), 503–513.

Zhang, L., Zhang, Y., & Zheng, X. (2020). Wisign: Ubiquitous american sign language recognition using commercial wi-fi devices. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *11*(3), 1–24.

Zhou, L., Meng, X., Huang, Y., Kang, K., Zhou, J., Chu, Y., Li, H., Xie, D., Zhang, J [Jiannan], Yang, W., & others (2022). An interpretable deep learning workflow for discovering subvisual abnormalities in CT scans of COVID-19 inpatients and survivors. *Nature Machine Intelligence*, *4*(5), 494–503.