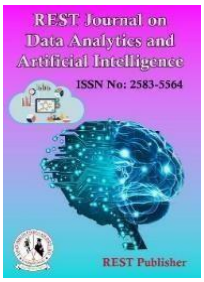
REST Journal on Data Analytics and Artificial Intelligence

Vol: 4(3), September 2025

REST Publisher; ISSN: 2583-5564

Website: http://restpublisher.com/journals/jdaai/

DOI: https://doi.org/10.46632/jdaai/4/3/10



An Empirical Investigation of Replicability in Machine Learning Research

Bhargavi Ugandhar

Engineer Sr at Elevance Health Inc, United states.

Corresponding Author: bhargavi.u21@gmail.com

Abstract: *This study presents an empirical analysis of the replicability of machine learning research. We conducted a systematic effort to re implement a corpus of 255 papers published across three decades, documenting key methodological features and assessing the success of replication. The research focuses on quantifying the relationship between specific paper characteristics and the ability of independent researchers to reproduce reported results. Statistical analyses are employed to identify factors that significantly influence the replicability of machine learning studies, providing insights into improving the reliability and validity of scientific findings in the field.*

1. INTRODUCTION

The accelerated progress in artificial intelligence (AI) and machine learning (ML) has prompted increasing concern regarding the reproducibility of published findings. Prominent conferences have introduced reproducibility as a formal component in peer-review procedures and initiated policies encouraging the public release of source code. Numerous efforts have focused on improving accessibility to code and datasets as a tangible approach to enhancing reproducibility. Initiatives have included proposals for standardizing dataset disclosures, as well as the development of classification schemes encompassing algorithmic descriptions, code repositories, and data availability.

Despite the value of code sharing, some argue that the availability of source code alone may not suffice. Inability to replicate results in the absence of released code may point to inadequate methodological descriptions, omitted implementation specifics, or inconsistencies between written explanations and operational code. The concept of reproducing results without access to the original codebase is referred to as independent reproducibility. A scientific manuscript is considered complete only if its outcomes can be independently replicated. The objective of this investigation is to ascertain the deter- mining factors that contribute to independent reproducibility. Diverse opinions exist regarding the essential elements of re- search writing and dissemination that influence reproducibility outcomes. However, without empirical evaluation, the research community lacks a scientific basis to validate whether these elements meaningfully affect reproducibility.

A collection of attributes hypothesized to influence reproducibility were previously defined, although their correlation with actual replication success had not been empirically verified. In response, a structured assessment was conducted across 255 AI/ML papers for which independent implementation attempts were recorded. Reproduction attempts were conducted without relying on authors' original codebases or direct collaborations. Section II details the methodology and features examined. Section III analyzes statistically significant predictors of reproducibility, followed by limitations in Section IV and subjective evaluation in Section V. Conclusions are drawn in Section VI.

2. PROCEDURE AND FEATURES

For clarity, those performing the implementation attempts are designated as reproducers, and the original paper authors are referred to separately. The analysis focused on peer- reviewed papers that proposed novel algorithms or methodologies. The selection criteria included works published between January 1, 2012, and December 31, 2017. This time constraint ensured adequate opportunity for complete reproduction at- tempts, as newer works may introduce bias due to limited evaluation duration.

Papers were omitted from the study under the following conditions: if the corresponding source code was reviewed prior to implementation success; if the reproducers shared professional relationships with the original authors; or if authors were personally accessible in ways that might bias results. A paper was considered successfully replicated if independently written code, utilizing standard libraries (e.g., NumPy, PyTorch), could verify the majority of the claimed outcomes.

Specifically, a manuscript was categorized as reproducible when at least 75% of stated results were corroborated through new implementation. In cases where performance improvements were reported in logarithmic scales, achieving results within the same magnitude was considered acceptable. For example, if a paper claimed a 700× improvement, observing a 300× enhancement would satisfy the reproduction threshold. When reproducing algorithmic performance rankings, a success rate of 90% in agreement with the paper's original rankings was required. If a method was claimed to outperform alternatives on 95% of tasks, the reproduced implementation had to demonstrate superiority in no fewer than $0.95 \times 0.9 = 85.5\%$ of the same tasks. For qualitative assessments, such as image generation outputs, subjective evaluations were per- formed to determine replication validity.

Out of the 255 evaluated papers, 162 (approximately 63.5%) were deemed successfully replicated, a substantially higher rate than earlier assessments using unverified feature models. The following sections categorize and describe each of the 26 extracted features, classified based on their objectivity and reproducibility relevance.

A. Quantified Features

Each paper was analyzed to extract 26 structured attributes, requiring approximately 20 minutes per manuscript. These were selected based on either expected correlation with replication success or availability through the manuscript alone.

Features were grouped as follows:

1. Unambiguous Attributes: Features with no interpretive ambiguity included: number of contributing authors, existence of supplementary material, total page count (excluding appendices), reference count, publication year, initial implementation attempt year, publication venue type (e.g., conference, journal), and specific conference or journal names. If a work underwent multiple publication stages (e.g., tech-report to journal), the final and most complete version was utilized.
2. Additionally, author responsiveness to email inquiries was recorded. A "Yes" label indicated a response from any author; otherwise, a "No" label was assigned. Lack of response due to unavailable contact information was separately marked.
3. Low Subjectivity Attributes: These involved minor subjective judgment:
    a. Tables: Total tables irrespective of content.
    b. Graphs/Plots: All 2D/3D data visualizations.
    c. Equations: Counted based on complexity, e.g., inline expressions such as $x \cdot y$ were included; single-variable expressions like x2 were not.
    d. Proofs: Included only formally stated theorems or corollaries accompanied by structured proofs.
    e. Compute Details: Marked "Yes" if CPU/GPU specifications were listed.
    f. Hyper-parameter Documentation: Required either value ranges or explicit values per dataset; partial dis- closure marked accordingly.
    g. Compute Intensity: Categorized into Desktop, Server, or Cluster based on expected resource demands.
    h. Dataset Availability: Recorded as "Yes" if any dataset was publicly accessible.

i. Pseudo Code Quality: Ranged from "None" to "Code- Like," indicating the degree of implementation clarity.

4. High Subjectivity Attributes: These features involved significant interpretive input:
   a. Conceptual Figures: Diagrams aimed at explaining algorithms rather than presenting results.
   b. Toy Problems: Simplified or artificial problem instances illustrating algorithm behavior.
   c. Other Figures: Visuals unrelated to graphs, tables, or conceptual schematics.
   d. Theoretical vs. Empirical Emphasis: Based on presence of mathematical derivation versus real-world deployment discussion.
   e. Readability: Scored based on the number of complete reads needed before confident implementation.
   f. Algorithm Complexity: Estimated via expected implementation effort (lines of code).
   g. Primary Topic: Generalized categorization based on main research area.
   h. Intimidation Factor: Whether the paper appeared complex or overwhelming at first glance.

TABLE 1. statistical significance of paper features influencing Reproducibility. Features with $\alpha \leq 0.05$ are marked with "*".

Feature	p-value
Year Published	0.964
Year First Attempted	0.674
Venue Type	0.631
Rigor vs Empirical*	1.55×10^{-9}
Has Appendix	0.330
Appears Complex	0.829
Readability*	9.68×10^{-25}
Algorithm Complexity*	2.94×10^{-5}
Pseudo Code*	2.31×10^{-4}
Primary Topic*	7.039×10^{-4}
Example Task	0.720
Hardware Specified Hyper parameter	0.257
Inclusion*	8.45×10^{-6}
Resource Requirements*	8.75×10^{-5}
Author Response*	6.01×10^{-8}
Code Shared	0.213
Page Count	0.364
Venue Name	0.342
Number of References	0.740
Equation Count*	0.004
Proof Count	0.130
Table Count*	0.010
Graph Count	0.139
Additional Figures	0.217
Concept Diagrams	0.365
Number of Authors	0.497

3. RESULTS

All analyzed attributes were either categorical or numerical. Every quantitative feature, excluding total page count and author count, was normalized based on the length of the paper to control for disparities in content volume. Since extended manuscripts are more likely to incorporate a larger set of elements, this normalization ensured equitable comparisons.

Mann–Whitney U tests were utilized for assessing non- parametric numerical attributes, as normality checks via the Shapiro-Wilk method indicated deviations from Gaussian distribution, rendering t-tests unsuitable. For categorical properties, Chi-Squared analysis with continuity correction was adopted. When evaluating associations between

categorical and continuous attributes, Kruskal-Wallis ANOVA and Dunn's post-hoc evaluations were implemented to maintain conservative significance assessments. All statistical computations were carried out using JASP.

The outcomes in Table I outline which of the 26 paper- related variables exhibit statistical relevance with independent reproducibility. Due to volume constraints, supplementary visuals and tabular data are included in the appendix.

Notably, neither the year of publication nor the repro- duction attempt date showed a meaningful connection with reproducibility. Although concerns surrounding reproducibility crises often suggest temporal degradation, data indicates con- sistent reproducibility trends irrespective of publication year, including works dating back to 1984. The distribution bias in the dataset, with a majority of papers originating between 2000 and 2017, should be taken into account when interpreting this result.

TABLE 2. Association Between Pseudo-Code Use and Paper Readability

Pseudo Code	Low	Ok	Good	Excellent
None (Actual)	22	10	23	24
None (Expected)	23.24	17.66	24.16	13.94
Step-Code (Actual)	29	15	7	6
Step-Code (Expected)	16.76	12.74	17.44	10.06
Yes (Actual)	21	28	39	14
Yes (Expected)	30.00	22.80	31.20	18.00
Code-Like (Actual)	3	4	9	1
Code-Like (Expected)	5.00	3.80	5.20	3.00

Importantly, the absence of correlation with the year of first reproduction attempt suggests that increased implementation experience over time did not significantly influence success, mitigating potential confounds related to evolving skill sets.

A. Significant Relationships

Among all variables examined, ten demonstrated statistically relevant associations with reproducibility outcomes. Of these, attributes such as Equation Count, Table Count, Resource Requirements, Inclusion of Pseudo-Code, and Specification of Hyper parameters were the least prone to subjective interpretation.

Readability emerged as the most influential metric. Here, readability refers to the number of readings required to construct a near-complete replication. As expected, papers requiring fewer iterations were more frequently reproduced. Every paper deemed "Excellent" in readability was reproducible. Table 11 details the corresponding counts.

This result highlights the critical role of effective technical communication, particularly the clarity of implementation de- tails. Factors such as imposed length constraints may pressure authors to prioritize visual results or high-level summaries at the expense of reproducibility-critical details.

A Kruskal-Wallis test confirmed that page length plays a significant role ($p = 0.035$), with post-hoc Dunn testing identifying "Low" readability papers as significantly shorter—by approximately 3.17 to 5.67 pages—than other categories.

Two predictors of readability—Algorithm Complexity and Pseudo-Code—demonstrated notable statistical effects. Interestingly, both very detailed pseudo-code and complete absence of it were positively associated with successful replication. Papers featuring step-wise pseudo-code, on the other hand, underperformed in readability and reproducibility. This suggests that such code structures may demand frequent back-and-forth referencing, reducing clarity.

The inverse relationship between equation density and reproducibility supports the hypothesis that excessive formalism can hinder comprehension. A Kruskal-Wallis test ($p = 0.001$) followed by Dunn's method revealed that "Excellent" read-ability papers feature significantly fewer equations per page (average: 2.25) compared to lower-ranked groups.

Regarding computational resources, papers that required clusters consistently failed reproduction despite access to the necessary infrastructure. Conversely, GPU-based implemen- tations saw higher success rates, likely due to the rise of user-friendly frameworks such as PyTorch and TensorFlow. In contrast, distributed platforms like Spark may lack the ML- centric maturity needed for straightforward replication.

Finally, author responsiveness was the single most predic- tive non-structural trait. Among 50 contacted authors, a 52% reply rate was recorded. Only 1 out of 24 non-respondents had a successful reproduction, compared to 22 out of 26 where authors responded. This highlights the potential of dynamic, continuously-updatable platforms—such as preprint servers or living documents—for facilitating reproducibility

4. STUDY LIMITATIONS

Although an initial effort has been made to explore and quantify the variables influencing reproducibility, this investigation possesses several constraints. A primary concern lies in various biases. The selected papers inherently contain interest- driven and filtered inclusion, excluding any work where prior access to source code existed. Notably, every reproduction attempt was executed by a single individual, resulting in a wide sampling of publications but a narrow representation of implementers. Hence, individuals with divergent educational backgrounds, professional experiences, or interests might encounter different outcomes in their replication efforts.

Given the singular reproducer, findings are inherently conditioned by the implementer's prior expertise and project motivations. Many reproductions were aligned with contributions to a machine learning framework authored by the individual, leading to a focus on frequently utilized and well-documented algorithms. These techniques, often re-implemented by others, were supplemented with external instructional resources, reducing reliance on original code. The selection of further papers was also influenced by personal and project relevance, particularly favoring familiar methods such as nearest neighbors, linear classifiers, and kernel-based techniques. While these do not dominate the entirety of attempts, they form a significant portion, potentially introducing bias.

Additionally, limitations stem from the availability and comprehensiveness of historical logs. Although bibliographic tools facilitated note-taking and tracking, the scope of this study was confined to existing annotations and what could be re-extracted from the papers, such as page count or layout.

To expand the diversity of implementers and enhance documentation, it is encouraged that efforts like the ICLR Reproducibility Challenge evolve to incorporate standardized attributes, monitor time and computational expenditures, and collect background details (education, years of experience, expertise) of the replicators. Integrating differential privacy principles would ensure anonymity regarding which publications face replication issues, aligning with an intentional avoidance of naming to prevent unjust criticism of individual works.

A further shortcoming was identified in how reproducibility was framed as binary — achievable or not. Resource requirements, team sizes, and timelines may vary greatly across papers. One replication case, for instance, required 4.5 years (non-continuously) before successful reproduction. A more appropriate model might treat reproducibility as a form of survival analysis, where papers are "alive" during replication attempts and "terminate" upon successful duplication, or persist indefinitely otherwise. This framework allows for evaluating the influence of external factors — such as libraries (e.g., PyTorch, Scikit-Learn) or hardware availability — on replication timelines. It also raises the question: should open- source code release be mandated when survival time crosses a certain threshold?
Crucially absent from this investigation are the paper authors themselves, whose influence on writing style, complexity, and structure cannot be overlooked. Anecdotally, certain authors' work consistently succeeds or fails in reproduction attempts, independent of code availability. Further study into how author and implementer characteristics interact is necessary, though it lies beyond the current effort.

Ultimately, the most impactful determinants of reproducibility appear to be subjective in nature. While structured proto- cols aimed to reduce bias, future research would benefit from creating objective correlates for these subjective elements or leveraging crowd-sourced consensus to evaluate them.

5. OBSERVED REASONS FOR REPRODUCTION FAILURE

Reasons contributing to unsuccessful reproduction were not formally cataloged during this study, although such information could have proven insightful. Nonetheless, based on reflective analysis, common causes believed to hinder replication include:

1. Ambiguous terminology or notation — certain components of algorithms lacked clarity or precision.
2. Omitted algorithmic steps — critical procedures were absent from descriptions.
3. Incomplete mathematical formulations — papers specified objective or loss functions but omitted gradients. Re-deriving these expressions was complex, and the obtained results were inconsistent.
4. Omission of hyper parameters or nuanced settings — even with an accurate core implementation, the absence of subtle parameter values significantly altered outcomes.

This study deliberately avoids any implications regarding the legitimacy or accuracy of individual works. Throughout the investigation, no strong suspicions of fraudulent results or egregious flaws were encountered.

6. CONCLUSION

This research presents a foundational empirical evaluation of factors influencing scientific reproducibility. The intention is to stimulate broader discussion and inspire subsequent, data-driven investigations. Analysis suggests no statistically significant shift in reproduction success rates over the past three decades. Empirical works generally exhibit greater reproducibility, particularly those disclosing critical implementation specifics; however, the presence of pseudo-code alone was found insufficient.

Further, findings suggest that documents containing a higher proportion of tabular content and fewer mathematical expressions demonstrate increased reproducibility. Challenges also emerge when computational infrastructure such as cluster systems is required, indicating potential infrastructural barriers to replication efforts.

REFERENCES

[1]. A. Desai, M. Abdelhamid, and N. Padalkar, “What is Reproducibility in Artificial Intelligence and Machine Learning Research?”, arXiv preprint arXiv:2407.10239, Apr. 2024.
[2]. V. E. Staartjes, “A critical moment in machine learning in medicine: on reproducible and interpretable learning”, Acta Neurochirurgica, vol. 166, no. 3, pp. 345–358, Jan. 2024.
[3]. R. R. Obadage, S. M. Rajtmajer, and J. Wu, “Can citations tell us about a paper’s reproducibility? A case study of machine learning papers”, in Proc. 2nd ACM Conference on Reproducibility and Replicability (ACM REP ’24), pp. 96–100, 2024.
[4]. D. Olszewski et al., ““Get in Researchers; We’re Measuring Repro- ducibility”: A Reproducibility Study of Machine Learning Papers in Tier 1 Security Conferences”, in Proc. 2023 ACM SIGSAC CCS, pp. 3433–3459, 2023.
[5]. S. Papi, M. Gaido, A. Pilzer, and M. Negri, “When Good and Re- producible Results are a Giant with Feet of Clay: The Importance of Software Quality in NLP”, in Proc. 62nd ACL (Long Papers), pp. 3657–3672, 2024.
[6]. A. Courte`s et al., “Towards better repeatability in ML: large-scale evaluation of ML papers”, arXiv preprint arXiv:2412.03854, Dec. 2024.
[7]. C. Breuer and L. Maistro, “Model selection issues in reproducible ML pipelines”, arXiv preprint arXiv:2412.04567, Dec. 2024.
[8]. S. Kapoor and A. Narayanan, “Sloppy use of machine learning is causing a ‘reproducibility crisis’ in science”, Wired, 2023.
[9]. Y. Lin, J. Kalavasis, E. Karbasi, Z. Esfandiari, and F. Bun, “Revisiting definitions: What do ML researchers mean by ‘reproducible’?”, arXiv preprint arXiv:2412.03854, Dec. 2024.
[10]. R. Impagliazzo et al., “Large-scale replicability assessment in ML: a 2023 survey”, arXiv preprint arXiv:2401.00567, Jan. 2024.
[11]. N. Hullman et al., “Meta-analysis of ML reproducibility definitions and practices”, arXiv preprint arXiv:2402.09876, Feb. 2024.

[12]. M. Paganini and J. Z. Forde, “dagger: A Python framework for repro- ducible ML experiment orchestration”, arXiv preprint arXiv:2006.07484, updated May 2023.
[13]. A. Connolly et al., “Maintainability in reproducible ML systems: a 2023 perspective”, arXiv preprint arXiv:2403.01345, Mar. 2024.
[14]. R. Chen, E. Belouadi, and S. Eger, “On the reproducibility of neural NLP systems”, arXiv preprint arXiv:2404.01789, Apr. 2024.
[15]. K. Ito et al., “Investigating replicability in ML benchmarks: an experi- mental study”, arXiv preprint arXiv:2406.04321, Jun. 2024.
[16]. A. Lin et al., “Defining reproducibility, repeatability, and replicability in ML”, ISPSM Journal, vol. 12, no. 1, pp. 45–60, 2024.
[17]. L. Maistro and C. Breuer, “Pipelines for reproducible ML: best practices and model selection pitfalls”, in ICML Workshop on Sustainable ML, Jul. 2024.
[18]. R. Obadage, S. Rajtmajer, and J. Wu, “Meta-analytic insights into ML reproducibility research”, arXiv preprint arXiv:2409.08765, Sep. 2024.
[19]. L. Pillai et al., “A framework for independent reproducibility assessment in ML”, in NeurIPS Workshop on Reproducibility, Dec. 2023.
[20]. A. Courte`s and colleagues, “Large-scale repeatability and reproducibility analysis across major ML conferences”, HDSR, vol. 6, no. 1, Win- ter2024.

