Journal of Advanced Robotics, Autonomous Systems and Human-Machine Interaction ISSN:3070-6491

PRIME OPEN ACCESS

Volume 2, Issue 2

Research Article

Date of Submission: 05 May, 2026
Date of Acceptance: 05 June, 2026
Date of Publication: 15 June, 2026

Improving Reproducibility in Machine Learning: Overview, Barriers, and Drivers

Thulani Myles Mzyece*

Independent Reasercher, USA

***Corresponding Author:** Thulani Myles Mzyece, Independent Reasercher, USA.

Citation: Mzyece, T. M. (2026). Improving Reproducibility in Machine Learning: Overview, Barriers, and Drivers. *J Adv Robot Auton Syst Hum Mach Interact, 2*(2), 01-06.

Abstract

Reproducibility remains a fundamental issue in machine learning (ML) research; in a replication study of highly cited AI papers, Gundersen et al. found that only half of the evaluable studies were reproducible to any extent [1]. This systematic review examined the multidimensionality of ML reproducibility across computational, statistical, and methodological dimensions. Important barriers as well as facilitators have been identified. The thematic analysis of peer-reviewed articles, combined with the technical documentation, allowed for the categorization of barriers into the following categories: technical barriers of nondeterminism and environmental instability; methodological barriers of poor reporting and the evaluation trap; and cultural barriers related to the incentive structure. The corresponding drivers included technical tooling ecosystems with containerization, experiment-tracking systems, standardization measures with reporting checklists, artifact-evaluation programmes, institutional interventions with educational integration, and policy requirements. These observations were summarized into an effective five-step model: protocol definition, environment specification, experiment tracking, validation protocols, and artifact preparation. The current research has shown that reproducibility enhancement is difficult to achieve without simultaneous technological, methodological, and institutional interventions.

Keywords: ML, Computational, Reproducibility, Validation, Empirical

Introduction

Reproducibility is a pillar of rigorous scientific inquiry, enabling research results to be independently verified and validated across computational disciplines. Semmelrock et al. (2025) defined reproducibility as the ability of independent investigators to draw the same conclusions from an experiment by following the documentation provided by the original investigators. Although the importance of modern ML is widely acknowledged, empirical studies of its practice have found evidence of a reproducibility crisis: only around half of published results can be replicated by independent researchers. This research provides a methodological review that integrates theoretical knowledge and an implementation plan across the multidimensional aspects of ML reproducibility. The current research highlights three important types of barriers: technical, methodological, and cultural that hinder the reproducibility efforts, and at the same time, examines tangible drivers such as standardized tooling ecosystems, policy interventions, and community-initiated efforts. The result is an effective five-phase framework that research teams can implement to increase reproducibility across the lifecycle of an experiment and, therefore, the empirical basis of ML science.

Background and Related Work

Reproducibility in Computational Science

Reproducibility concerns were well established in the computational research community long before modern machine learning, with decades of evolution in other areas of scientific computing. With the integration of computing with the experimental and statistical sciences, the question is whether the classical principles of reproducibility could be applied to computational science [2]. Besides, the frameworks are developed concurrently with computing infrastructure and have moved beyond deterministic batch processing systems to distributed, heterogeneous computational environments that introduce new scales of complexity. Early areas of scientific computing established strict standards for numerical accuracy, algorithm definition, and result verification that later informed reproducibility models.

Existing ML Reproducibility Frameworks

Modern ML studies have introduced various frameworks that attempt to define reproducibility principles by providing systematic instructions and empirical measurement procedures. Early taxonomies that distinguish computational reproducibility from statistical reproducibility clarify that rerunning identical code does not guarantee that results remain stable. Several venues now require checklists on reproducibility, where hyperparameters, random seeds, computational resources, and the version of the data used must be explicitly recorded and included with the manuscript. Empirical research on the reproducibility of significant ML conferences yields grim findings: the likelihood of reproducing results depends significantly on task complexity and publication perimeter [3]. Such studies indicate that the code's availability alone is insufficient, as artifact execution often fails due to undocumented dependencies and environmental assumptions. Current frameworks have significant weaknesses, such as excessive implementation burden and insufficient coverage of statistical enforcement approaches beyond voluntary compliance. The lack of alignment between aspirational guidelines and the researcher's working process suggests that existing frameworks remain underutilized in real development processes. New artifact appraisal initiatives in major institutions aim to fill these gaps through systematic review procedures, but uptake is recent, and rates are low.

Adjacent Research Streams

Similar work has also been conducted that concentrates on specific issues related to the reproducibility problem as a way of complementing those technological and methodological efforts. According to Whelan and Patterson (2025), reproducibility remains imperative for innovation but also quite challenging to implement. Model cards and dataset datasheets are documentation structures that help standardize the communication of key metadata, enabling downstream users to understand the training conditions and potential limitations. Model cards are short documents that accompany trained machine learning models and provide benchmarked evaluations across conditions relevant to the intended application domains, including cultural, demographic, phenotypic, and intersectional groups. The transparency frameworks provide machine-readable details that facilitate the automatic verification of compatibility in research work. Moreover, reproducibility has been improved through the use of experiment tracking systems such as MLflow, Weights and Biases, and Neptune, which automatically record configurations, statistics, and artifacts. MLflow is a widely used open-source platform for managing ML development, including experiment tracking, reproducibility, and deployment [4,5]. The research software engineering methodology, which has been adopted by other disciplines, primarily emphasizes the importance of version control and containerization as infrastructure that fosters reproducibility. The systems minimize manual documentation and enable systematic comparison of results across different experimental runs, although they are less prevalent in research fields. Ahmed et al. (2025) analyzed the concept of reproducibility in deep learning driven biodiversity studies, revealing that few articles passed complete reproducibility tests.

Methodology

The current research used a systematic narrative review to summarize the available literature and identify common trends across reproducibility studies in machine learning. Systematic searches were performed in IEEE Xplore, ACM Digital Library, Google Scholar, and arXiv databases. The use of search terms related to machine learning, reproducibility, and replicability, using Boolean operators AND/OR. Only peer-reviewed articles in the English language published in the time frame between 2015 and 2025 were considered. The preliminary search yielded 247 articles; the retrieved articles were then reviewed for title and abstract (198 articles). The inclusion criteria demanded empirical reproducibility studies, theoretical frameworks, or specific documentation of technical tools. A total of 45 papers were reviewed in full, and 16 were eligible for final inclusion. The review is based on primary sources, including official framework documentation, technical specifications, and guidelines from key research venues and standards bodies. Secondary materials include peer-reviewed empirical research on quantifying reproducibility, technical surveys on impediments to reproducibility, and analyses of reproducibility challenges presented at major conferences. Demirezen et al. (2024) developed research principles on reproducible ML using electroencephalography data and demonstrated the effectiveness of systematic literature reviews in professional circles. In their study, Demirezen et al. (2024) used thematic synthesis to identify common patterns of barriers across the technical, methodological, and cultural aspects, creating a taxonomy in which drivers are categorized into tooling ecosystems, standardization initiatives, and policy interventions.

Dimensions of ML Reproducibility

The concept of ML reproducibility has three dimensions that are interconnected, each with distinct challenges and specific solutions for verifying and validating its presence. Computational reproducibility means that the bitwise same results are obtained when the same code is run on the same hardware setup using the same input data and software requirements. Desai et al. (2024) have explained the terminology of validation by differentiating between repeatability, dependent and independent reproducibility, direct and conceptual replicability in the artificial intelligence setting. The main conditions are a detailed environment description through containerization, clear random-seed control across all stochastic elements, and records of platform-specific behaviors that may influence numerical results. The particularly challenging aspects of ML are nondeterministic operations on the GPU that are sensitive to the order of parallel computation, incompatibilities between framework versions, and hardware dependencies that lead to divergent behavior. Statistical reproducibility goes beyond computational exactness to include the statistical stability of results across repeated experimental runs with controlled randomization, accounting for natural variability in stochastic optimization processes. In this dimension, it is important to report accurate estimates of variance, confidence intervals, and significance tests across multiple independent runs with varying random initializations. Based on the findings of

Desai et al. (2024), in sixteen equal training runs of LeNet5, accuracy ranged from 8.6% to 99% across the same hyperparameters and random seeds. Vincent and P. (2022) proposed a better hyperparameter optimization framework based on evolutionary algorithms that demonstrates statistical significance, as evidenced by the Wilcoxon rank-sum and Friedman tests. Besides, hyperparameter optimization adds complexity that needs to be documented in search strategies and computational budgets, and in selection criteria to differentiate incidental performance improvements from hard improvements. Method reproducibility addresses whether independent applications of methods described in published reports produce similar results, challenging the adequacy and clarity of methodological reporting. Code availability and code clarity are important distinctions, as opaque implementations can succeed without revealing key insights into an algorithm. In 2020, the Association for Computing Machinery proposed a badging system to certify the research artifact at various levels, ranging from Available and Evaluated to Reproduced.

Barriers to Reproducibility

Technical and Computational Barriers

Environmental instability is a key technical barrier that manifests as dependency conflicts, version drift, and platform-specific behavioral differences across computational infrastructure. Software ecosystems are constantly evolving, with new framework releases introducing breaking changes, deprecating libraries, removing formerly available functionality, and operating system changes that alter runtime behavior. Container adoption studies show that Docker and Singularity significantly increase the reproduction success rate by enabling the installation of entire software environments. Sources of non-determinism are widespread in current ML pipelines, affecting both the algorithmic and hardware levels. The randomness of algorithms includes schemes to initialize weights, shuffle data, generate dropout masks, and sample stochastic gradients, which influence the optimization path. The issue of hardware-level nondeterminism arises from the parallel operation of a set of GPUs, where the order of thread execution affects floating-point accumulation, resulting in numerically distinct outcomes per execution. The PyTorch documentation notes that deterministic operations are often slower than nondeterministic ones, and that the performance of single runs may be reduced when deterministic algorithms are used [6]. Quantified case studies measure these effects and show performance differences of more than 5 % points that can only be explained by nondeterministic execution under the same hardware configurations. Data management challenges contribute to reproducibility failures in various ways, including data leakage, lack of versioning, and privacy limitations that restrict sharing. Kapoor and Narayanan (2023) conducted a systematic study of data leakage across 17 scientific areas and found that it occurred in 294 studies, leading to wildly overoptimistic scientific conclusions in ML. A lack of data versioning does not effectively monitor the evolution of datasets, because corrections to annotations, the addition of new examples, or the optimization of collection procedures can occur over time.

Methodological Barriers

Incomplete reporting is a widespread methodological impediment in which publications lack essential details of the implementation needed to reproduce an experiment faithfully. The details of hyperparameter search are often not reported, hiding the vast amount of tuning work and leaving an immense effect on the final performance, but it is not visible to readers. Semmelrock et al. (2025) identified three primary issues that prevent reproduction: specified or underspecified ML models or training processes, specified evaluation metrics, and selective reporting of results. These omissions form the largest portion of reproduction failures, as evidenced by reproducibility challenge postmortems.

The pitfalls of evaluation compromise the validity of results due to train-test contamination, overuse of benchmarks, and insufficient statistical characterization of performance variability. Poor cross-validation methods can introduce information leakage, where the performance of the validation set is used to inform model selection and, ultimately, the validation information appears in the final test set. In a paper by Ahmed et al. (2025), the methodology of assessing the reproducibility of publications related to biodiversity using deep learning was developed, with results revealing that not all publications provided adequate randomness control and a statistical perspective. Even publications accompanied by code repositories have documentation gaps because their availability does not ensure that independent researchers can understand or execute the code.

Cultural and Institutional Barriers

Incentive misalignment is a core problem of reproducibility. Existing academic reward systems encourage novelty over rigor in the choice of publication venues and in career progression. The value given to publication velocity carries more weight than methodological thoroughness, which creates time pressure, discourages the extra effort required for comprehensive documentation and artifact preparation, and undermines the value of these efforts. Little credit is given to researchers who invest in reproducibility infrastructure. Resource constraints are practical because extensive hyperparameter search requires significant computational resources and is often impossible for researchers outside well-funded laboratories. Competitive research situations put time pressure on researchers, compelling them to prioritize quick publication over proper documentation. The insufficient institutional support for reproducibility infrastructure means that a single researcher will have to cover the costs of artifact preparation and documentation programmes. Besides, there is a lack of standardized educational materials, where reproducibility training is an informal practice specific to the lab and not taught directly.

Drivers of Reproducibility Improvement

Technical Tooling Ecosystem

Environment management tools, especially containerization systems such as Docker and Singularity, provide dependency

isolation by encapsulating entire software stacks, including operating systems, libraries, and framework versions. It has been empirically proven that containerized artifacts have a significantly higher success rate in reproduction than traditional installation instructions. Numerous ML-specific images are now hosted in container registries, making duplication of such images easier. Experiment tracking tools, such as MLflow, Weights and Biases, and OpenML, formalise run data management by logically recording parameters, data, artifacts, and system configurations. Chen et al. (2020) reported changes to MLflow, an open platform for managing the machine learning lifecycle, experimentation, reproducibility, deployment, and a central model registry. These platforms remove the manual record-keeping overhead and permit systematic cross-comparison of thousands of experimental runs. Frameworks such as PyTorch and TensorFlow include determinism controls that allow execution of otherwise stochastic training procedures to be reproducible. Seed management utilities allow all random number generators used for weight initialization, data shuffling, dropout sampling, and augmentation to be initialized consistently. Besides, deterministic algorithms are implemented at the expense of speed by framework-specific reproducibility modes.

Standardization and Best Practices

Reporting standards exemplified by the NeurIPS reproducibility checklist have had a statistically significant effect on submission quality, as empirically evaluated, and code submission rates have risen significantly after the checklist's introduction. Mitchell et al. (2019) proposed model cards as a standard for reporting on models, which comprise formalized documentation on the use of ML models, potential restrictions, and ethics. Public recognition mechanisms, such as ACM's badging system, which assesses the availability, evaluation, and reproduction of artifacts, promote reproducibility. SIGMOD 2008 was the first database conference to offer to test submitters' programs on their data to verify the experiments published [7]. Besides, FAIR principles: Findable, Accessible, Interoperable, Reusable, modified to the data management environments, offer an organizing scheme of the practices of sharing the artifacts of ML. Ravi et al. (2022) proposed FAIR principles for artificial intelligence models and provided practical implementation in accelerated high-energy diffraction microscopy. Huerta et al. (2023) considered FAIR as an AI concept within the context of international, interdisciplinary community building, stating that making data and AI models FAIR helps one better understand their content and context.

Policy and Institutional Interventions

The increasing venue requirements at large-scale conferences include submitting code with publications, and some venues require executable artifacts that have been reviewed before admission. NeurIPS, ICML, and other major conferences now have reproducibility tracks, which make replication studies more visible and legitimate. The development of policies over time shows that the requirements are being strengthened gradually. Demirezen et al. (2024) developed guidelines for reproducible EEG-based machine learning research, organized according to the Cross Industry Standard Process for Data Mining. Educational integration entails integrating reproducibility training into ML program curricula through systematic workflow training for student research teams, rigorous capstone project requirements, and direct instruction in best practices for version control, containerization, and documentation.

Figure 1: Five-Stage Workflow with Barriers and Drivers

Practical Framework

Synthesizing insights from barrier and driver analyses yields an actionable five-stage workflow that research teams can implement throughout experimental lifecycles. Protocol definition: stage one focuses on prior registration of research questions, hypotheses, and evaluation metrics before access to data, a priori of post hoc rationalization. Prospective data split specifications avert leakage by documentary training, validation, and test partitions. Resource constraints are established in advance in the budget documentation of the computation. Stage two, environment specification, uses container-based environment capture and documents the entire software stack, including base operating systems, framework versions, and auxiliary libraries. Lock files make specific packages fixed. Documentation of hardware and platforms captures GPU architecture, driver version, and CUDA toolkits. Experiment tracking is the third stage, which uses automated logging of every configuration, including hyperparameters, random seeds, dataset versions, and preprocessing parameters. The Hyperparameter search strategy documentation captures the optimization strategies and search spaces that have been tried. Result variance is characterized by multiple run executions with controlled seeds. The fourth stage, the validation protocol, involves conducting statistical tests of significance between runs to identify improvements that are not due to chance. Ablation experiments separate single-component effects. Overfitting is avoided by performing independent validation on the held-out data. The fifth stage, artifact preparation, produces small, reproducible illustrations that show core contributions. FAIR-compliant documentation includes detailed README files, installation documentation, and usage examples. Before submission, peer review and code review identify errors.

Discussion and Conclusion

Although reproducibility is an essential requirement of rigorous research, it is not enough on its own to guarantee research quality. Full reproducibility imposes real strain on resources due to computational costs. The fundamental obstacles are privacy and proprietary data restrictions. Domain-related issues must have specific requirements regarding reproducibility. These must consider the practical constraints. The hosting venues must have substantial, not nominal, assessments of artifacts, and the institutions of the respective professionals must provide training and infrastructure so that assistance with reproducibility can be offered. In conclusion, to enhance the reproducibility of ML, interventions to improve reproducibility must be aligned with technical tooling, methodological practices, and institutional incentive mechanisms that work together, not individually. The systematic review of the existing research identifies a range of practical obstacles, including environmental instability, the inability to report methods, and conflicting career motives. Standard drivers, such as containerization tools, standardized checklists, and policy requirements, show plausible directions to increased reproducibility. These insights are combined into a five-stage framework for practical advice. Reproducibility serves not as bureaucratic compliance but as fundamental infrastructure that enables cumulative scientific progress [8-16].

Conflicts of Interest

I declare that I have no conflicts of interest.

References

1. Gundersen, O. E., Cappelen, O., Mølnå, M., & Nilsen, N. G. (2024). The unreasonable effectiveness of open science in AI: A replication study
2. Lewis, J., Breeze, C. E., Charlesworth, J., Maclaren, O. J., & Cooper, J. (2016). Where next for the reproducibility agenda in computational biology? BMC Systems Biology, 10(1).
3. Ugandhar, B. (2025). An empirical investigation of replicability in Machine Learning Research. REST Journal on Data Analytics and Artificial Intelligence, 4(3 September 2025), 73–79.
4. 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. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229.
5. Chen, A., Chow, A., Davidson, A., DCunha, A., Ghodsi, A., Hong, S. A., ... & Zumar, C. (2020, June). Developments in mlflow: A system to accelerate the machine learning lifecycle. In Proceedings of the fourth international workshop on data management for end-to-end machine learning (pp. 1-4).
6. PyTorch Contributors. (2024). Reproducibility. PyTorch Documentation.
7. Manolescu, I., Afanasiev, L., Arion, A., Dittrich, J., Manegold, S., Polyzotis, N., ... & Shasha, D. (2008). The repeatability experiment of SIGMOD 2008. ACM SIGMOD Record, 37(1), 39-45.
8. Ahmed, W., Kommineni, V. K., König-Ries, B., Gaikwad, J., Gadelha, L., & Samuel, S. (2025). Evaluating the method reproducibility of deep learning models in biodiversity research. PeerJ Computer Science, 11, e2618.
9. Demirezen, G., Taşkaya Temizel, T., & Brouwer, A. M. (2024). Reproducible machine learning research in mental workload classification using EEG. Frontiers in Neuroergonomics, 5, 1346794..
10. Demirezen, G., Taşkaya Temizel, T., & Brouwer, A. M. (2024). Reproducible machine learning research in mental workload classification using EEG. Frontiers in Neuroergonomics, 5, 1346794..
11. Desai, A., Abdelhamid, M., & Padalkar, N. R. (2025). What is reproducibility in artificial intelligence and machine learning research?. AI Magazine, 46(2), e70004.
12. Huerta, E. A., Blaiszik, B., Brinson, L. C., Bouchard, K. E., Diaz, D., Doglioni, C., ... & Zhu, R. (2023). FAIR for AI: An interdisciplinary and international community building perspective. Scientific data, 10(1), 487.
13. Ravi, N., Chaturvedi, P., Huerta, E. A., Liu, Z., Chard, R., Scourtas, A., ... & Foster, I. (2022). FAIR principles for AI

models with a practical application for accelerated high energy diffraction microscopy. Scientific Data, 9(1), 657.
14. Semmelrock, H., Ross-Hellauer, T., Kopeinik, S., Theiler, D., Haberl, A., Thalmann, S., & Kowald, D. (2025). Reproducibility in machine-learning-based research: Overview, barriers, and drivers. AI Magazine, 46(2), e70002.
15. Vincent, A. M., & P., J. (2022). An Improved Hyperparameter Optimization Framework for AUTOML Systems Using Evolutionary Algorithms.
16. Whelan, M., & Patterson, E. (2025). Achieving reproducibility in the innovation process. Open Research Europe, 5, 25.